REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Ecole Supérieure en Sciences Biologique d'Oran (ESSB d'Oran)
Département des Classes Préparatoire ou Département du Second Cycle

Polycopié Pédagogique

Matière

BIOSTATISTIQUES

-Cours-

Réalisé par :

Dr. NASSER Soraya

Niveau : 2ème année des classes préparatoires
Filière : Sciences Biologiques
Domaine : Science de la Nature et de la Vie

Année Universitaire : *2022/2023*

TABLE DES MATIÈRES :

1. INTRODUCTION :

La statistique est une branche des mathématiques appliquées qui implique la collecte, la description, l'analyse et l'inférence de conclusions à partir de données quantitatives, elle s'est développée à partir de l'application d'outils mathématiques, y compris le calcul et l'algèbre linéaire à la théorie des probabilités.

Les statistiques sont utilisées dans pratiquement toutes les disciplines scientifiques telles que les sciences physiques et sociales, ainsi que dans les affaires, les sciences humaines.

Les deux grands domaines de la statistique sont les statistiques descriptives et la statistique inférentielles.

. De ce fait, la science de la Biostatistique englobe à la fois : (1) La conception des expériences biologiques ; (2) La collecte des informations ; (3) La compilation et analyse des données chiffrées de ces expériences ; (4) L'interprétation des résultats en vue d'avancer une conclusion. Cette science est exploitée alors dans plusieurs domaines biologique: (1) La santé publique, y compris l'épidémiologie, les services de santé, la nutrition et l'environnement ; (2) La conception et analyse d'essais cliniques en médecine ; (3) La génomique, génétique des populations et la génétique statistique afin de relier la variation dans le génotype avec une variation dans le phénotype ; (4) L'agriculture afin d'améliorer les cultures et les animaux d'élevage ; (5) L'écologie en vue de mettre en place des prévisions écologiques ; (6) L'analyse de séquences biologiques.

Cette discipline constitue, en biologie, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence aux biologistes comme :

- Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?
- Quelle est la fiabilité d'une mesure ou d'une observation ?
- Quel est le risque ou l'avantage d'un traitement ?
- Les conditions expérimentales A sont-elles plus efficaces que celles des conditions de B ?
- Les effets de la variable A sont-ils les mêmes ou diffèrent-ils des effets de la variable B ?

La Biostatistique est une matières destinées aux étudiants des classes préparatoires et de spécialités de l'école. Le but recherché par ces programmes d'enseignement est d'initier l'étudiant aux concepts et aux méthodes statistiques :

- Introduire l'étudiant aux mesures statistiques principales.

- L’interprétation, et l’utilisation de distributions de données.
- Introduire les concepts de probabilité, probabilité conjointe, probabilité conditionnelle et leurs relations.
- Introduire le concept d’étude d’observation, d’échantillonnage et d’inférence statistique de l’échantillon.

A la fin du cours l’étudiant sera capable de reconnaître les mérites et les faiblesses de la méthode statistique, Il sera capable de réaliser des opérations quantitatives simples comme l’interprétation d’une valeur « anormale » ou le calcul de la probabilité. Il interprétera correctement le concept de test statistique.

Le polycopié est constitué de plusieurs chapitres, dont les trois premiers chapitres sont consacrés à la statistique descriptive, uni-variée et Bi-variée, en expliquant les différents concepts adoptés dans la démarche statistique, suivi d’un détail sur la nature des caractères étudiées et les différentes techniques d’organisation et de structuration d’une série de données. Chaque section est accompagnée d’un ensemble d’exemples explicatifs.

A partir du quatrième chapitre en entame la statistique inferentielle par une introduction aux calculs de probabilités contenant l’analyse combinatoire, les propriétés des probabilités, les variables aléatoires discrètes et continues puis on passe vers les la distribution d’échantillonnage, la théorie d’estimation et les tests statistiques.

2. CHAPITRE 1 : STATISTIQUES DESCRIPTIVES À UNE DIMENSION

2.1. Définition :

La statistique descriptive est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données. Il est assez compliqué de définir la meilleure description possible d'un phénomène. Dans le cadre des statistiques, il s'agira de fournir toute l'information disponible sur le phénomène en moins de chiffres et de mots possibles.

2.2. Notions de bases

2.2.1. Elément : C'est une unité qui peut être :

Un individu (êtres vivants) : humain, animal, végétal ... ;

Un sujet : modules enseignés en biologie, les nationalités, les métiers ou professions… ;

Un objet : Table, chaise, verrerie de laboratoire ;

Une association : dans les études écologiques en général : une parcelle d'herbe, une association d'arbustes…

2.2.2. Population : C'est un ensemble d'éléments possédant au moins une caractéristique commune et exclusive permettant de l'identifier et la distinguer sans ambiguïté de toutes les autres.

Exemple : Une population algérienne ; Une population estudiantine ; Une population de plantes médicinales ; Une population de poissons d'eau douce.

2.2.3. Echantillon

Pour des raisons techniques ou économiques, il n'est généralement pas possible de collecter des données sur tous les éléments de la population. En outre, si cette opération est possible il est rarement utile de la faire, car l'analyse d'un groupe restreint d'éléments extraits de la population fournit généralement des résultats de précision satisfaisante. Cette petite partie de la population qu'on va examiner s'appelle « échantillon ».

Exemple :

- Etude de 20 étudiants pris à partir d'une population de 57.
- Etude de 5 régions prises à partir d'une population de 25

2.2.4. Echantillonnage :

C'est l'opération ou la méthode qui consiste à prélever une partie de la population (échantillon). Il existe plusieurs méthodes d'échantillonnage qui varient en fonction de la nature de l'étude envisagée. Le plus utilisé est l'échantillonnage aléatoire et simple qui est basé sur le principe que tous les éléments de la population ont une probabilité égale (non nulle) de faire partie de l'échantillon. C'est une méthode d'échantillonnage permettant de choisir n unités parmi les N de la population de façon aléatoire.

2.2.5. Caractère :

Chaque étude portera sur un ou plusieurs caractères présentés par chacun des individus de la population

Exemple : taille, poids, taux de glycémie, couleur des yeux, profession, nationalité……etc.

Un caractère est dit quantitatif quand ses différentes modalités sont mesurables (exprimées par des chiffres).

Exemple : taille, poids, nombre d'enfants par famille

Un caractère est dit qualitatif quand ses différentes modalités échappent à la mesure

Exemple : couleur des yeux, profession, groupe sanguin…etc.

Le caractère qualitatif peut être

Nominal : le sexe, le groupe sanguin.

Ordinal : la douleur, le niveau d'instruction.

2.2.6. Modalité :

Ce sont les diverses situations (cas, état, valeur) susceptibles d'être prises par le caractère. Un caractère peut posséder une ou plusieurs modalités.

Exemple :

- Couleur des yeux : vert, bleu, noir ;
- Poids des souris (en grammes) : 15, 18, 20, 39 ;
- Superficie d'une pièce (en mètres) : 3, 5, 6 ;
- La température de l'air (en °C) : 8, 16, 27, 30, 38.

2.2.6.1. Effectif : L'effectif de la valeur xi est le nombre d'individus de la population ayant cette valeur ou appartenant à cette classe : on le note $\mathbf{n_i}$.

L'effectif total N est la somme de tous les effectifs, appelé également en Maths « cardinal ou la taille ». En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer l'effectif cumulé croissant en faisant la somme des effectifs de cette valeur et de tous ceux qui la précèdent.

Exemple

Dans un groupe de 20 étudiants, voici les notes obtenues au dernier examen de Biostatistique : 10, 14, 12, 15, 7, 8, 10, 11, 12, 18, 2, 4, 12, 13, 14, 15, 19, 11, 9, 0.

On va calculer les effectifs et les effectifs cumulés.

Tableau I

Notes x_i	0	2	4	7	8	9	10	11	12	13	14	15	18	19
Effectif n_i	1	1	1	1	1	1	2	2	3	1	2	2	1	1

Les effectifs cumulés maintenant. On fait la somme des effectifs de la note + la somme des effectifs de toutes les notes qui la précédent. Ce qui nous donne (Tableau II):

Tableau II : Les effectifs ni et Ni↑ relatifs à chaque modalité xi (notes)

Notes x_i	0	2	4	7	8	9	10	11	12	13	14	15	18	19
Effectif n_i	1	1	1	1	1	1	2	2	3	1	2	2	1	1
Effectifs cumulés Ni↑	1	2	3	4	5	6	8	10	13	14	16	18	19	20

2.2.6.2. Fréquence d'une modalité ou d'une classe

La fréquence d'une valeur est le quotient de l'effectif de la valeur par l'effectif total

En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer les fréquences cumulées croissantes en faisant la somme des fréquences de cette valeur et de tous ceux qui la précèdent ;

Pour les fréquences cumulées croissantes, c'est un peu le même principe que pour les effectifs cumulés croissants.

Remarque : Les fréquences sont comprises entre 0 et 1.

Exemple **:**
On reprend l'exemple précédent et on applique tout simplement la formule des fréquences pour les calculer (Tableau III).

Tableau III : Les fréquences fi et Fi↑ relatives à chaque modalité xi (notes).

Notes x_i	0	2	4	7	8	9	10	11	12	13	14	15	18	19
Fréquences f_i	0,05	0,05	0,05	0,05	0,05	0,05	0,1	0,1	0,15	0,05	0,1	0,1	0,05	0 ,05
Fréquences $F_i\uparrow$	0,05	0,1	0,15	0,2	0,25	0,3	0,4	0,5	0,65	0,7	0,8	0,9	0,95	1

2.3. Nature de caractère

2.3.1. Caractère qualitatif

Un caractère est dit qualitatif lorsque ses modalités ne sont pas mesurables. Le nombre de valeurs que peut prendre la variable est limité. Il existe au sein de ce type deux échelles : nominale et ordinale.

- **Echelle nominale :** Chaque modalité est exprimée par un nom ou un code. Les différentes modalités ne sont pas ordonnables.

  Exemple :

- Etat matrimoniale : marié, célibataire, veuf, divorcé ;
- Sexe : féminin, masculin ;
- Profession : enseignant, médecin ;
- Nationalité : Algérienne, Tunisienne ;

Cas des codes

- **Echelle ordinale :** Chaque modalité est explicitement significative du rang pris par chaque individu pour le caractère considéré.

  Exemple :

- Degré d'intelligence : pas intelligent (0), peu intelligent (1), moyennement intelligent (2), très intelligent (3) ;
- Forme des fruits : petite (1), moyenne (2), grosse (3) ;

2.3.2. Caractère quantitatif

Un caractère est quantitatif si ses modalités s'expriment par des nombres. Le nombre de valeurs que peut prendre la variable est illimité.

Exemple **:** Latitude, longitude, température, altitude

Il peut être discret ou continu

Enfin, on peut distinguer les caractères quantitatifs discrets et les caractères quantitatifs continus selon que leurs modalités (valeurs) sont définies sur un intervalle continu de l'ensemble des réels (modalités en nombre infini) ou selon qu'elles correspondent à un ensemble fini et dénombrable de valeurs entières ou réelles.

a. Les caractères quantitatifs discrets : Sont des caractères dont les modalités sont des nombres isolés, pas nécessairement entiers.

Exemple :

- Nombre de pièces d'un immeuble ;
- Nombre d'enfants d'une famille ;
- Nombre des doigts

b. Les caractères quantitatifs continus : Sont des caractères dont les modalités sont définies sur un intervalle (continu) de valeur donné appelé domaine de variation et défini par les valeurs minimales et maximales.

Exemple **:** Notes des étudiants, la taille, le poids, l'âge**.**

2.4. Tableaux statistiques

Un tableau statistique constitue un résumé ou une synthèse numérique des résultats d'une distribution statistique, on distingue trois formes de tableaux statistiques qui sont fonction de l'objectif envisagé et de la nature du caractère étudié.

2.4.1. Tableaux élémentaire (brut) : Ce type de tableau permet de recueillir l'ensemble des données brutes obtenues dans le cadre d'une étude.

Elément i	Modalité xi
1	X1
2	X2
3	X3
:	:
K	Xk

2.4.2. Tableau de dénombrement

Il est de la forme :

Classes	Centre c_i	Effectifs n_i	Fréquences relatives f_i
$[e_{i-1}, e_{i-1} + a_i[$	c_1	n_1	f_1
$[e_{i-1} + a_i, e_{i-1} + 2a_i[$	c_2	n_2	f_2
$[e_{i-1} + 2a_i, e_{i-1} + 3a_i[$	c_3	n_3	f_3
.	.	.	.
.	.	.	.
.	.	.	.
	c_k	n_k	f_k
Total	-	n	1

Exemples

a. Tableau statistique relatif à un caractère qualitatif

Analyse de sang pour 100 étudiants

Groupe sanguin xi	ni	fi
A	40	O.4
B	43	0.43
AB	12	0.12
O	5	0.05
Total	100	1

b. Tableau statistique relatif à un caractère quantitatif

- Cas discret

On observe 20 lots au laboratoire, on a le nombre de lapins dans chacun : 10, 1, 0, 0, 0, 10, 12, 18, 5, 5, 12, 10, 12, 12, 0, 10, 15, 10, 20, 20

x_i	0	1	5	10	12	15	18	20	T
n_i	4	1	2	5	4	1	1	2	20
fi	0.20	0.05	0.10	0.25	0.20	0.05	0.05	0.10	1

Nombre de classes

Il existe plusieurs formules pour le calcul du nombre de classes. Pour les applications de ce cours, nous avons opté pour la Règle de STURGE - La formule mathématique de HUNTSBERGER

K=1+3.3 log N

Intervalle de classe (amplitude)

$$a_i = \frac{x_{\max} - x_{\min}}{k}$$

Centre d'une classe

$$c_i = \frac{e_{i-1} + e_i}{2}$$

Cas continu

On s'intéresse à la taille (cm) de 20 étudiants, les résultats obtenus sont : La taille de 20 étudiants (cm) :**140 ,142 ,143 ,143,144,144,146,147,148,150,150,152,153,154,155,156,157,158,159,163 .**

Nombre de classes k=5,29 ~5 classes ; Amplitude ai= 4,6 ~ 5

Tableau de la distribution des tailles des étudiants :

Classes	ci	ni	fi
[140-145[	142.5	6	0.30
[145-150[	147.5	3	0.15
[150-155[	152.5	5	0.25
[155-160[	157.5	5	0.25
[160-165[	162.5	1	0.05
Total	-	20	1

2.5. Représentation graphique

2.5.1. Cas qualitatif

2.5.1.1. Diagramme en tuyaux d'orgue

Le diagramme en bâtonnets (ou tuyaux d'orgue) est une représentation graphique de la distribution de fréquences d'une variable qualitative. Les « bâtonnets» sont bien séparés pour indiquer les différentes catégories. La hauteur d'un bâtonnet est proportionnelle à la fréquence de la catégorie correspondante.

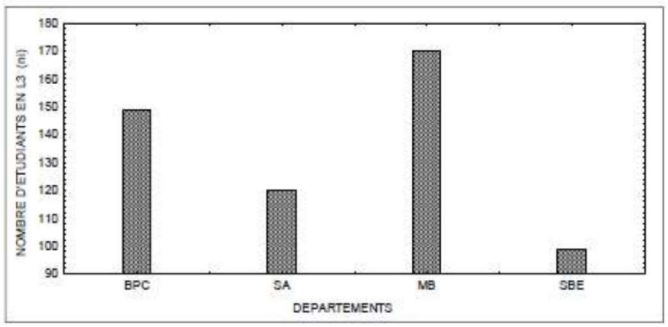


Figure 1 : Tuyaux d'orgue illustrant le nombre d'étudiants en L3 par département (**SA** : Science Alimentaire, **MB** : Microbiologie, **SBE** : Science Biologique de l'Environnement, **BPC** : Biologie Physico-Chimique).

2.5.1.2. Le camembert

Dans le diagramme circulaire, chaque secteur a une surface proportionnelle à la fréquence de chaque modalité (Figure 2).

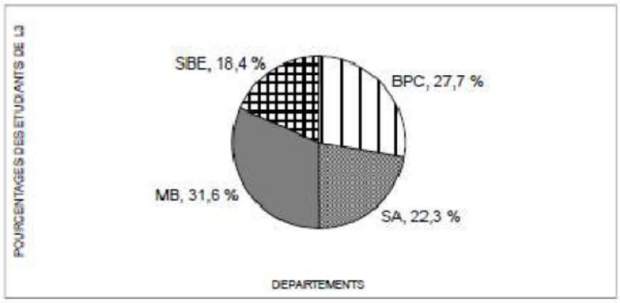


Figure 2 : Le camembert illustrant le nombre d'étudiant en L3 en fonction des départements (**SA** : Science Alimentaire, **MB** : Microbiologie, **SBE** : Science Biologique de l'Environnement, **BPC** : Biologie Physico-Chimique).

2.5.2. Cas quantitatif

Il existe deux types de représentation graphique d'une distribution statistique à caractère quantitatif :

- Le diagramme différentiel correspond à une représentation des effectifs ou des fréquences.
- Le diagramme intégral correspond à une représentation des effectifs cumulés, ou des fréquences cumulées.

2.5.2.1 Variable statistique discrète.

1. Diagramme différentiel (Diagramme en bâtons) : est réalisé en fonction des effectifs ou des fréquences.

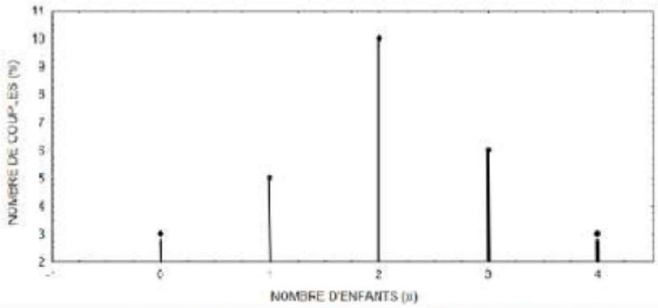


Figure 3 : Diagramme différentiel illustrant le nombre de couples en fonction du nombre d'enfants.

---

2. Diagramme intégral (diagramme en Escalier)

Courbe en escalier : est réalisée en fonction des effectifs cumulés ou des fréquences cumulées. Dans cette représentation les effectifs ou les fréquences des diverses valeurs de la variable statistique correspondent aux hauteurs des marches de la courbe

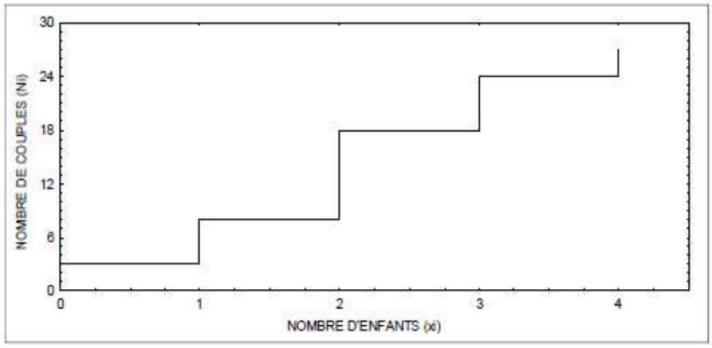


Figure 4 : Diagramme intégral illustrant le nombre de couples en fonction du nombre d'enfants.

2.5.2.2 Variable statistique continue

1. Histogramme et polygone des effectifs ou des fréquences

L'histogramme est une représentation graphique (en tuyaux d'orgue) de la distribution des effectifs ou des fréquences d'une variable quantitative. Souvent, les «tuyaux» sont accolés pour montrer la continuité de la variable. La hauteur du tuyau est proportionnelle à l'effectif ou la fréquence de la classe correspondante (Figure 5).

Le polygone des effectifs ou des fréquences : est une autre représentation graphique (en ligne brisée) de la distribution des effectifs ou des fréquences d'une variable quantitative.

Pour tracer le polygone, on joint les points milieu du sommet des rectangles adjacents par un segment de droite. Le polygone est fermé aux deux bouts en le prolongeant sur l'axe horizontal.

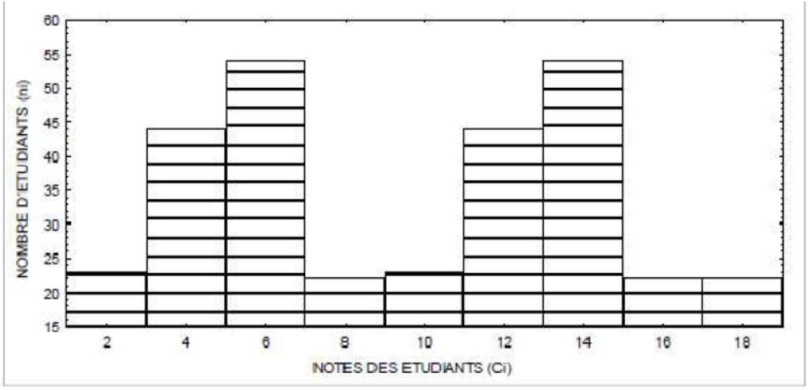


Figure 5 : Histogramme illustrant le nombre d'étudiants en fonction de leurs notes.

2. Diagramme intégral

Représentation graphique intégrale en fonction des effectifs cumulés ou des fréquences cumulées appelée parfois ogive (Figure 6). Une telle figure fournit des données descriptives intéressantes telles que la valeur médiane,

i.e. le point séparant le groupe en deux parties égales.

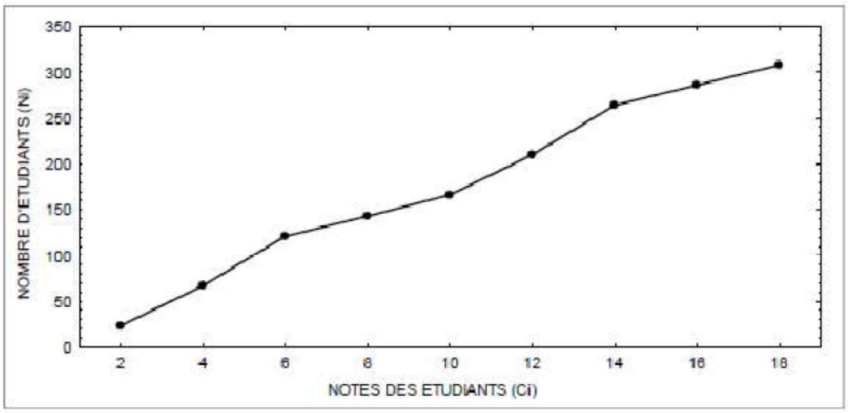


Figure 6 : Diagramme intégral illustrant le nombre d'étudiants en fonction de leurs notes.

2.6 Description numérique :

L'objectif de la description numérique est de résumer en un seul nombre la valeur typique. En statistiques, les distributions de variables sont caractérisées à travers trois critères, qui suffisent généralement : Forme de la distribution, tendance centrale et dispersion.

2.6.1Tendance centrale ou Paramètres de position

1. **Le mode**

Le mode est la valeur la plus fréquente d'une distribution. Il se calcule toujours à partir d'un dénombrement des modalités du caractère.

a. Caractère qualitatif

- Le mode d'une variable qualitative est la modalité la plus fréquemment observée
- Le mode est la seule mesure de la seule tendance centrale applicable aux variables qualitatives

b. Caractère quantitative discret :

Pour un caractère quantitatif discret le mode est la modalité qui a la fréquence la plus élevée (ou l'effectif le plus élevé).

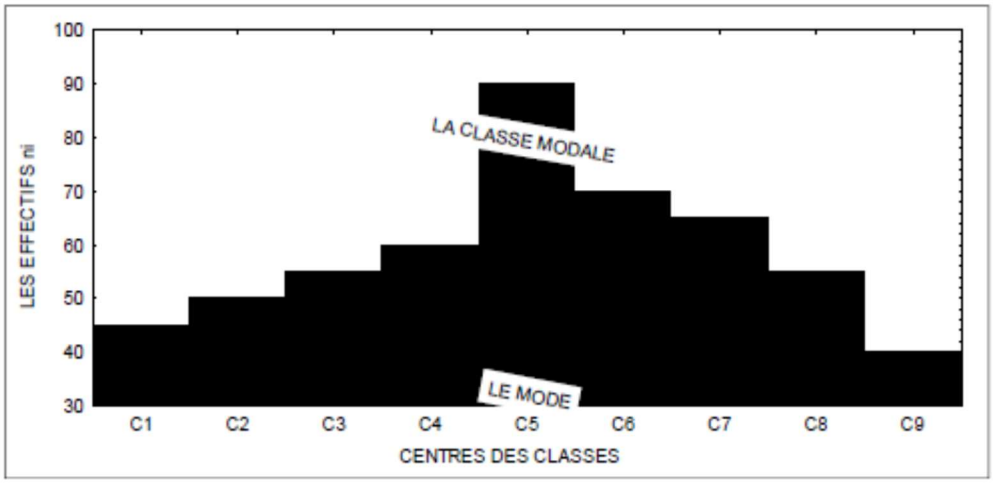
Exemple :

- 10, 11, 12, 10, 10, 10, 9, 14 → Mode : 10 (4 fois) → Distribution unimodale.
- 10, 11, 12, 10, 10, 12, 12, 9, 14 → Modes : 10 (3 fois) et 12 (3 fois) → Distribution bimodale.

c. Caractère quantitatif continue :

a. Méthode directe :(Approche1)

Il faut au préalable établir une partition en classes. Le mode est alors **le centre de la classe modale**, c'est à dire la classe qui a l'effectif le plus élevé ou la fréquence la plus élevée.



b. Méthode indirecte : (Approche 2)

Dans ce cas on peut calculer le mode par une deuxième méthode appelée « **le calcul approché** », en utilisant la formule :

$$Mo = e_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2}.$$

Δ_1 : C'est l'excès de la classe modale par rapport à la classe précédente.

Δ_2 : C'est l'excès de la classe modale par rapport à la classe suivante.

e_{i-1} : C'est la borne inférieure de la classe modale.

a_i : C'est l'amplitude

La figure 8 explique les différents paramètres de cette formule.

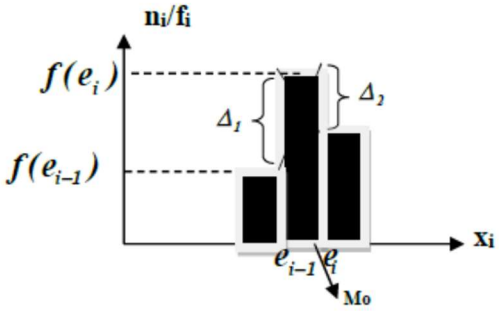


Figure 8 : Histogramme précisant le calcul du mode par la 2ème approche.

2. **La médiane**

Les valeurs étant classées par ordre croissant, la médiane est la valeur du caractère qui partage La série statistique en deux ensembles d'effectifs égaux : 50 % des valeurs lui sont supérieures et 50 % lui sont inférieures.

a. **Caractère quantitatif discret**

N est impaire → N=2p+1

$Me = x_{(p+1)}$

p : est l'ordre de la variable dans la série statistique supposée ordonnée dans l'ordre croissant ou décroissant.

N est pair → N = 2p

$Me= [X_{(p)} +X_{(p+1)}]/2$

Exemple :

- ✓ 0, 0, 0, 0, 1, 5,5, 10, 10, 10, 10, 10, 12, 12, 12, 12, 15, 15, 18, 20, 20.
- ✓ n=20 (pair) : n=2*p → p=20/2=10.
- ✓ Me=[x10+x11] / 2 → Me= [10+10] / 2=10.

a. **Caractère quantitatif continue**

Calcul par la méthode d'interpolation linéaire

$$\mathbf{Me}=e_{i-1}+\frac{\frac{N}{2}-S}{N_{me}}\,a_i$$

e_{i-1} : la borne inferieure de la classe médiane

N : la taille de la distribution statistique

S :les effectifs qui précèdent la classe médiane

Nme : l'effectif de la classe médiane

ai :l'amplitude des classes

Remarque : la classe médiane est la classe dont l'effectif cumulé est supérieur ou égale à N/2

Ou bien la sa fréquence est supérieure ou égale à 0.5

3. **Les quantiles**

La notion de **quantile** (ou de **fractile**) est le terme général pour désigner les valeurs qui divisent un jeu de données en intervalles contenant le même nombre de données.

a- **Les quartiles :** On appelle **quartile** les modalités d'une variable quantitative qui divise la population en 4 parties égales, soit en 3 quartiles (Q1, Q2, Q3).
Un quart (25 %) des valeurs sont inférieures au premier quartile Q1.
Le deuxième quartile Q2 est aussi la médiane puisqu'il sépare l'effectif en deux parties égales.
Un quart des valeurs (25 %) sont supérieures au troisième (et donc dernier) quartile Q3.

b- **Les deciles : sont les** valeurs qui partagent une distribution en dix parties égales, chacune contenant 10 % de l'effectif**.** Il y a 9 déciles, notés D1 jusqu'à D9

c- **Les quintiles :** séparent la population en 5 parties égales de 20 % (un cinquième de l'effectif). Il y a 4 quintiles notés V1 jusqu'à V4 .

d- **Les centiles :** Les centiles partagent l'étendue des valeurs en cent sous- ensembles d'effectifs égaux, chacune comprenant 1 % de l'effectif total, notés C1 jusqu'à C99.

4. **La moyenne**

C'est un critère de position. Il s'agit d'une mesure statistique caractérisant les éléments d'un ensemble de quantités et elle exprime la somme de toutes les mesures divisées par l'effectif total de l'échantillon étudié.

Cas quantitatif discret :

$$\bar{x} = \frac{\sum n_i x_i}{N}$$

x_i : les observations (modalités)

n_i : l'effectif correspond à l'observation x_i

N : la taille de la distribution statistique

Cas quantitatif continu :

Dans ce cas on utilise les mêmes formule que dans la cas précédent sauf que les modalités seront les centres de classes.

$$\bar{x} = \frac{\sum n_i c_i}{N}$$

2.6.2 Comparaison des valeurs centrales

Il n'y a pas de règle générale pour déterminer laquelle des mesures de tendance centrale est la plus pertinente pour caractériser une distribution.

Exemple *:*
Les nombres ci-dessous représentent le nombre d'enfants dans chacune de dix familles choisies au hasard :3, 8, 1, 1, 4, 5, 2, 3, 0, 1 (il s'agit d'étude d'un échantillon).

- ✓ **Moyenne**

$$m = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}}{n} = 1/10\ [(3+8+1+1+4+5+2+3+0+1)] = 28$$

10

m= 2,8

- ✓ **Médiane**

Rangeons les données en ordre ascendant :
0, 1, 1, 1, 2, 3, 3, 4, 5, 8
Comme n = 10 (donc pair), p = 5
Me = ($5^{ème}$ Observation + $6^{ème}$ Observation) / 2 = (2 + 3) / 2 = 2,5

- ✓ **Mode**

On voit immédiatement que *Mo* = 1
On remarque que les trois mesures sont différentes.
Laquelle est la plus «représentative» ? (Voir ce qui suit.)

2.6.3 Valeurs centrales et forme des distributions

La comparaison des trois valeurs centrales ou l'examen de l'histogramme permettent de définir la forme des distributions. Selon la forme d'une distribution le meilleur résumé sera fourni par l'une ou l'autre des trois valeurs centrales.

a. Distributions bimodales ou multimodales

La distribution comporte plusieurs modes. Le mode principal est différent de la moyenne et de la médiane. Dans ce cas, ni la moyenne ni la médiane ne sont significatives.
Le meilleur résumé est alors donné par les modes principaux et secondaires de la distribution.

b. Distributions unimodales symétriques

Lorsque la distribution est unimodale et symétrique, on va trouver à peu près
moyenne =médiane = mode.
Le meilleur résumé est alors donné par la moyenne car elle tient compte de toutes les observations et elle possède des propriétés statistiques intéressantes.

c. Distributions unimodales dissymétriques

Deux cas peuvent se présenter :

- ✓ mode < médiane < moyenne : la distribution est dissymétrique à gauche (ou la distribution est positivement dissymétrique (Figure 10 A). c'est à dire qu'il y a concentration pour les valeurs faibles et dispersion pour les valeurs fortes. C'est le cas le plus fréquent
- ✓ moyenne < médiane <mode : la distribution est dissymétrique à droite (ou la distribution est négativement dissymétrique (voir la distribution B), c'est à dire qu'il y a concentration pour les valeurs élevées et dispersion pour les valeurs faibles. Ce cas est plus rare que le précédent.

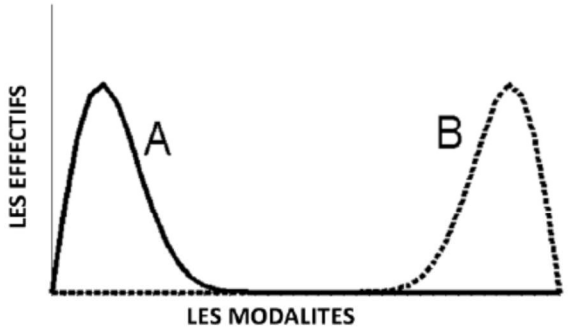


Figure 10 : Courbe illustrant les deux types de distribution. A : ditribution dissymétrique à gauche. B : distribution dissymétrique à droite.

Remarque :

Lorsqu'il y a une dissymétrie marquée, la médiane est généralement préférable à la moyenne car elle est moins influencée par les valeurs exceptionnelles qui sont souvent à l'origine de la dissymétrie. Si la dissymétrie est peu marquée, on pourra néanmoins utiliser la moyenne.

2.6.4 les paramètres de dispersion

Quantifient les fluctuations des valeurs observées et leur étalement

1.Variance : La variance est l'écart carré moyen entre chaque donnée et le centre de la distribution représenté par la moyenne.

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

ou bien la formule suivante

$$V = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \overline{x}^2,$$

On peut le nomé S

2. L'ecart-type : $\mathbf{S = \sqrt{S^2}}$

3. La fourchette des valeurs ou étendue

L'étendue E est la différence entre la plus grande valeur et la plus petite valeur. La formule générale est : $\mathbf{E=V_{max}-V_{min}}$

4. L'intervalle interquartile :

(Q_3-Q_1) est un paramètre de dispersion absolue qui correspond à l'étendue de la distribution une fois que l'on a retiré les 25% des valeurs les plus faibles et les 25% des valeurs les plus fortes. 50% des observations sont donc concentrées entre Q_1 et Q_3.

$\mathbf{I_q= Q_3-Q_1}$

5. Coefficient de variation

Ce coefficient correspond à l'écart type de la distribution exprimé en pourcentage de la moyenne de la distribution

La formule générale est :

$$CV = \frac{S}{\overline{x}} * 100$$

Ce nombre est sans unité, c'est une des raisons pour lesquelles il est parfois préféré à l'écart type. En effet, pour comparer deux séries de données d'unités différentes, l'utilisation du coefficient de variation est plus judicieuse.

2.6.5 Paramètres de dispersion absolue et valeurs centrales

1. Le résumé d'une distribution par une valeur centrale est souvent trompeur parce qu'il est incomplet : on connaît l'ordre de grandeur des valeurs mais on ignore la dispersion des valeurs autour de la valeur centrale de référence. Le premier rôle des paramètres de dispersion est donc :

D'accompagner et de préciser les résumés de distribution effectués à l'aide des valeurs centrales.

2. Un bon résumé statistique doit donc toujours comporter au moins deux paramètres : une valeur centrale et un paramètre de dispersion. L'appariement des valeurs centrales et des paramètres de dispersion ne peut toutefois pas s'effectuer de n'importe quelle manière et certaines associations peuvent se faire de façon privilégiée.

La moyenne peut-être accompagnée de l'écart-type (puisque ces paramètres sont calculés par rapport à la moyenne).
La médiane peut-être accompagnée soit de l'intervalle interquartile
Le mode, qui n'est véritablement utile que dans le cas des distributions multimodales, peut être accompagné de l'étendue

3. CHAPITRE 2 : STATISTIQUES DESCRIPTIVES À DEUX DIMENSIONS

3.1 PRESENTATION D'UNE SERIE A DEUX VARIABLES

L'objectif de cette étude statistique est d'étudier sur une même population de N individus, deux caractères différents (ou modalités différentes) et de rechercher s'il existe un lien ou une Corrélation entre ces deux variables.

Exemple de relations possibles entre les variables suivantes : taille et âge ; diabète et poids ; taux de cholestérol et régime alimentaire ; niche écologique et population ; ensoleillement et croissance végétale ; toxine et réaction métabolique ; survie et pollution ; effets et doses; organe 1 et 2 ; organe et fonction biologique ; …

Les caractères étudiés peuvent être aussi bien qualitatifs que quantitatifs.

Les résultats sont généralement représentés sous forme d'un **tableau à double entrée**, appelé **tableau à deux dimensions**, ou **tableau croisé** ou **tableau de contingence**, ou parfois **tableau de corrélation**.

Exemple de tableau de contingence

	Effets de doses (variable y)			
Sexe (variable x)	**Effet 1**	**Effet 2**	**Effet 3**	**total**
H	43	36	3	**Total des H** : 82
F	49	12	12	**Total des F** : 73
Total	**Total effet 1 :** 92	**Total effet 2 :** 48	**Total effet 3 :** 15	**Total des H et F** : 155

Effets de doses selon le sexe H ou F

3.2 Représentations graphique

3.2.1 Diagrammes pour deux variables qualitatives

On effectue pour chacune des modalités d'une des variables un diagramme représentant l'autre variable. Les figures ci-dessous nous fournissent deux exemples. Dans le cas où l'une des variables est ordinale, il convient de tenir compte de l'ordre des modalités dans la représentation.

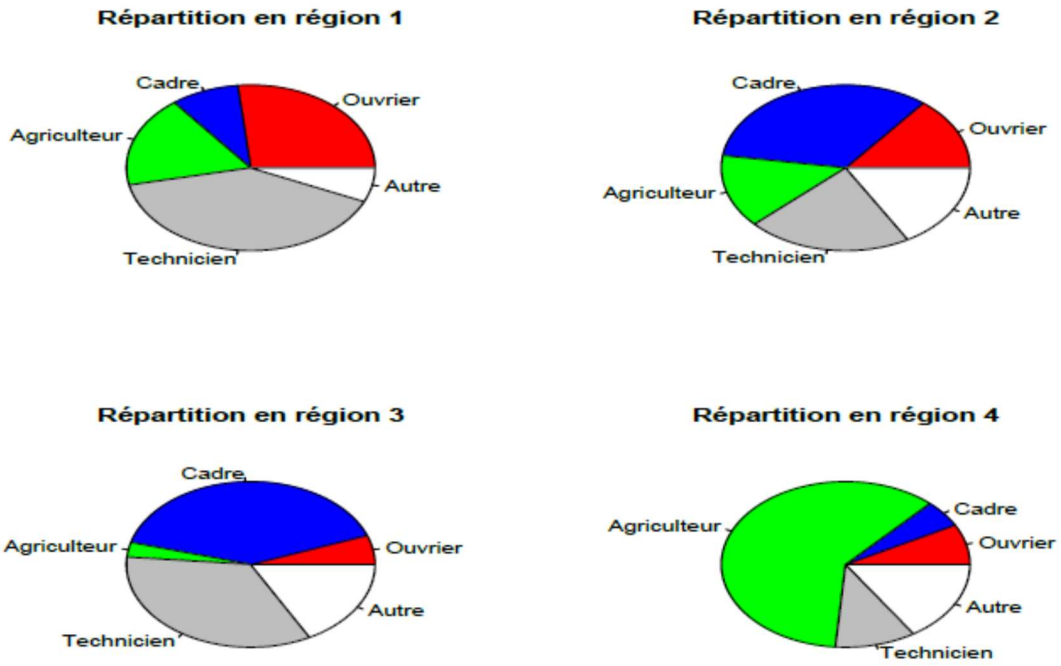


Figure : Exemple de représentation de 2 variables qualitatives.

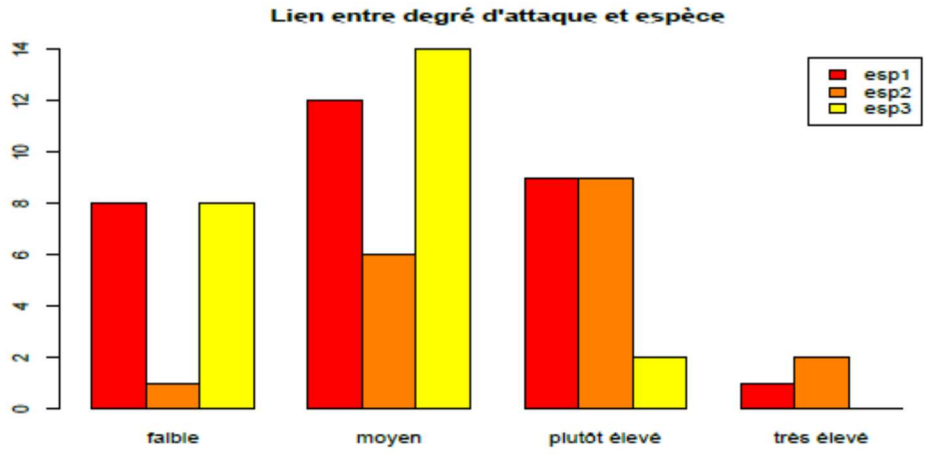


Figure : Exemple de 2 variables qualitatives, une nominale et l'autre ordinale.

3.2.2 Diagrammes pour cas mixte

Quand une variable est qualitative et l'autre est quantitative, le diagramme le plus approprié est souvent le box-plot (aussi appelé diagramme boitte à moustaches), à raison d'un box-plot par modalité de la variable qualitative.

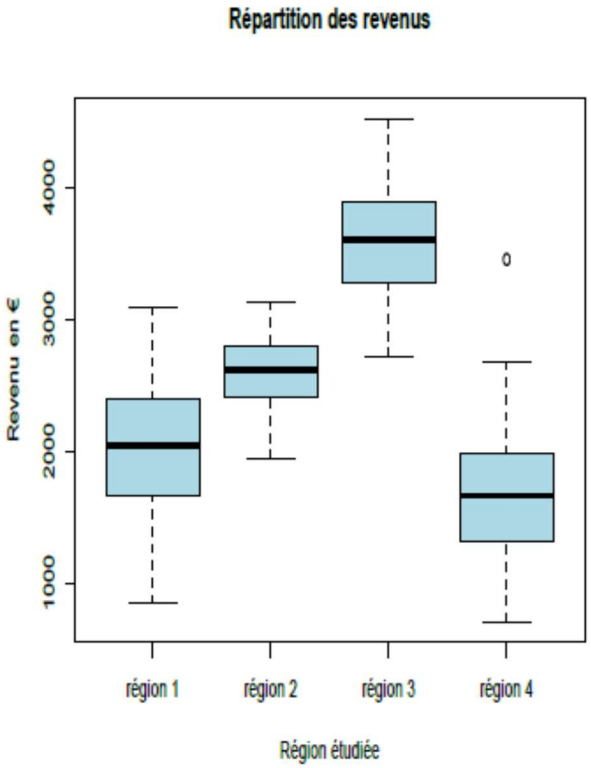


Figure : exemple de représentation d'une variable quantitative et une variable qualitative continue

Ce diagramme est constitué d'une boite dont la première arête est positionnée en Q1 et la seconde en Q3 et de deux moustaches de longueur au plus égale à **1,5*(Q3-Q1).**

La boite symbolise 50% des valeurs (les valeurs en centre de distribution). La position de la médiane Me, séparant la boite en deux, permet de visualiser l'éventuel étalement des valeurs.

Les valeurs inferieures à **Q1-1,5*(Q3-Q1)** ou supérieures à **Q3+1,5*(Q3-Q1)** sont dites atypiques car trop éloignées des valeurs centrales.

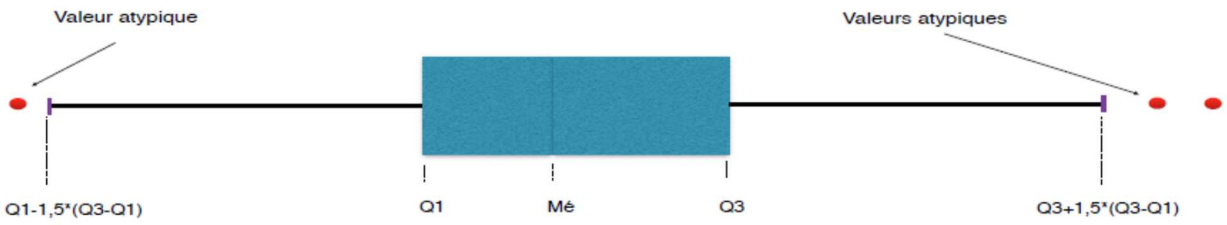


Figure : Boite à moustaches.

On peut également effectuer un autre type de diagramme pour chaque modalité de la variable qualitative (histogramme, diagramme en bâtons,...) selon que la variable quantitative est discrète ou continue.

3.2.3 Diagrammes pour deux variables quantitatives

Quand les deux variables sont quantitatives, on utilise la représentation en nuage de points : chaque individu statistique est représenté dans le plan par un point de coordonnées égales aux valeurs observées sur cet individu. La proximité de deux points dans le plan correspond à la similarité des couples de valeurs associées aux deux variables statistiques. Le nuage peut avoir une allure particulière qui nous renseigne sur le lien éventuel entre les deux variables (allure rectiligne(Droite), exponentielle, parabolique, etc)

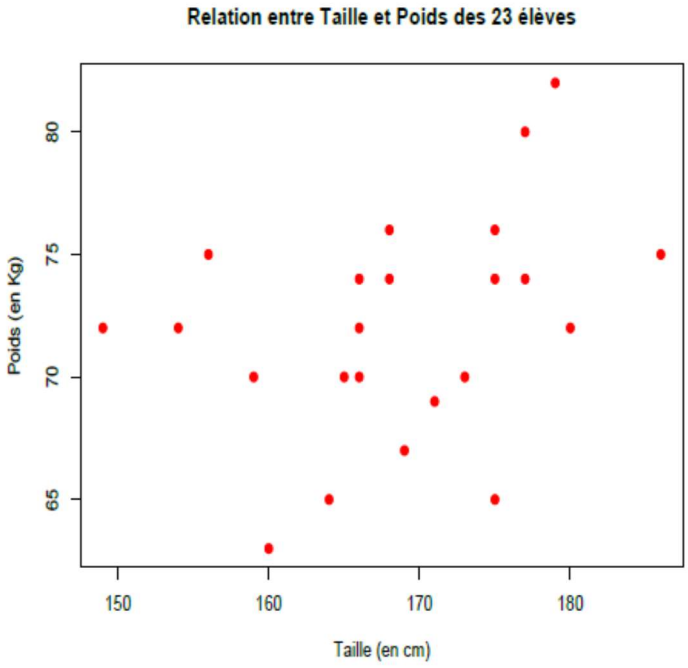


Figure : Nuage de points

3.3 Tableaux de contingence

Soient Z et T deux variables qualitatives à modalités respectivement $z_1,\ldots\ldots z_k$ et $t_1,\ldots\ldots..t_l$
Les valeurs de ces variables ont été observées sur une population de n individus. La répartition des effectifs suivant les modalités de Z et de T, se présente sous forme d'un tableau à double entrée, appelé tableau de contingence ou encore tableau croisé :

$Z \setminus T$	t_1	$\cdots$	t_j	$\cdots$	t_l	Total
z_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet l}$	n

L'**effectif nij** qui se trouve sur la i-ème ligne et la j-ème colonne est le **nombre d'individus qui possèdent à la fois la modalité zi de la variable Z et la modalité tj de la variable T**.
Ces effectifs ; n_{ij} : $i = 1, ..., k$ $j = 1, ..., l$ sont appelés les **effectifs croisés observés**.

L'**effectif ni•** qui se trouve sur la i-ème ligne et la **colonne Total** est le nombre d'individus qui possèdent la modalité zi de la variable Z ; on a donc

$$n_{i\bullet} = n_{i1} + n_{i2} + \ldots + n_{il}$$

L'**effectif n•j** qui se trouve sur la j-ème colonne et la **ligne Total** est le nombre d'individus qui possèdent la modalité tj de la variable T ; on a donc

$$n_{\bullet j} = n_{1j} + n_{2j} + \ldots + n_{kj}$$

L'effectif n qui se trouve sur la ligne Total et la colonne Total est le nombre d'individus de la population étudiée ; on a donc

$$n = n_{1\bullet} + n_{2\bullet} + \ldots + n_{k\bullet} \quad \text{et} \quad n = n_{\bullet 1} + n_{\bullet 2} + \ldots + n_{\bullet l}\,.$$

La fréquence de la modalité zi de la variable Z est donnée par : $f_{i\bullet} = \dfrac{n_{i\bullet}}{n}$.

La fréquence de la modalité tj de la variable T est donnée par: $f_{\bullet j} = \dfrac{n_{\bullet j}}{n}$.

Pour les moyennes : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\, x_i$ et $\overline{y} = \dfrac{1}{n}\sum_{j=1}^{l} n_{\bullet j}\, y_j$

Pour les variances : $Var(X) = \dfrac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\, x_i^2 - \overline{x}^2$ $\quad Var(Y) = \dfrac{1}{n}\sum_{i=1}^{k} n_{\bullet j}\, y_j^2 - \overline{y}^2$

3.4 Résumés numériques d'une série bi variée

Quand notre série est bivariée, on peut pour chacune des deux variables statistiques mener séparément une étude univariée. Aux résumés numériques vus dans le cas d'une série univariée viennent se rajouter des résumés concernant les liaisons éventuelles entre les deux variables.

3.4.1 Cas de deux variables quantitatives

L'objectif de l'analyse bi-variée est d'étudier les éventuelles relations entre deux variables statistiques

A- Covariance de deux variables X et Y

Soient x1; x2; ... ; xN et y1; y2; ... ; yN les valeurs prises par X et Y pour une population de N individus. La covariance de X et Y notée par cov(X; Y) est définie par

$$\text{cov}(X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_N - \bar{x})(y_N - \bar{y})}{N},$$

où $\bar{x}$ et $\bar{y}$ désignent les moyennes arithmétiques de X et Y ; notons que :

cov(X; X) = Var(X) et cov(X; Y) = cov(Y; X)

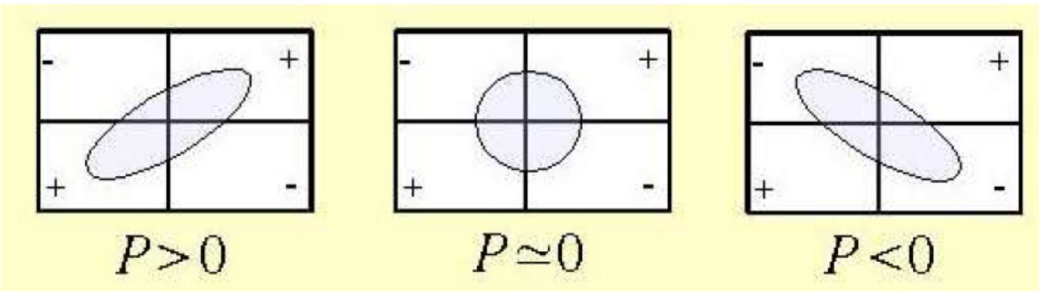
La covariance de X et Y peut aussi être calculée au moyen de la formule (parfois désignée par **formule de Huygens**) :

$$\text{cov}(X, Y) = \left(\frac{x_1 y_1 + x_2 y_2 + \ldots + x_N y_N}{N}\right) - \bar{x}\bar{y};$$

$$\text{COV(X,Y)} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{i,j}(x_i - \bar{x})(y_j - \bar{y})$$

Propriétés de la covariance :

- Cov(x,y)= Cov(y,x)
- Cov(x,x) = Var(x)
- Le signe de la Cov est un indicateur de la tendance de la relation sens positif ou négatif (direction d'étirement du nuage de point)
-



Une covariance positive indique une tendance « croissante » des valeurs de Y en fonction de

X, une covariance négative une tendance « décroissante »

B- Coefficient de corrélation linéaire de deux variables X et Y

$$r_{x,y} = \frac{Cov(x,y)}{S_x S_y}$$

Propriétés importantes du coefficient de corrélation linéaire

1) R(X; Y) est toujours compris entre -1 et +1
2) Lorsque **R(X; Y) est voisin de 0**, pas de liaison linéaire, mais possibilité d'une liaison d'un autre type.
3) Lorsque **R(X; Y) est voisin de +1**, il y a une **corrélation directe** entre les variable X et Y ; cela signifie que Y augmente lorsque X augmente et inversement.
4) Lorsque **R(X; Y) est voisin de -1**, il y a une **corrélation inverse** entre les variables X et Y ; cela signifie que Y augmente lorsque X diminue et inversement.

C- DROITE DE REGRESSION OU D'AJUSTEMENT

Une droite de régression linéaire s'écrit selon l'équation : « **y = ax +b** ». Cette approche de corrélation repose sur l'hypothèse que la relation entre deux variables est de nature linéaire.
En faite, il est possible de soupçonner une relation différente entre ces variables :

- courbe exponentielle
- courbe logarithmique,
- courbe hyperbolique, etc…

En partant de l'équation $y = ax + b$, a et b doivent être choisis convenablement de sorte que la droite passe au plus proche (ou par le plus possible) des points expérimentaux. Pour ce faire, on utilise la méthode des moindres carrés : On cherche les coefficients a et b de la droite qui minimise la somme des carrés des distances entre les points expérimentaux et la droite de régression (les points théoriques).

$$a = \frac{Cov(x,y)}{var(x)}$$

$$b = \bar{y} - a\bar{x}$$

4. CHAPITRE 3 : CALCUL DES PROBABILITÉS

4.1. Analyse combinatoire :

C'est un ensemble de méthodes et de techniques qui consistent à choisir, énumérer des objets, à dénombrer les différentes manières de classement, de groupement des éléments dans un ou plusieurs ensembles. Cette partie des mathématiques est appelée également « dénombrement ». Elle est largement utilisée en probabilité et en statistique en particulier dans le dénombrement des cas favorables et les cas possibles du rapport classique d'une probabilité, dans le binôme de Newton, le triangle de Pascal, la loi binomiale…

4.1.1 Arrangements (ordre) :

Etant donné un ensemble (**E**) de (***n)*** objets, on appelle arrangements de (**p**) objets **toutes suites ordonnées** de (**p**) objets pris parmi les (**n**) objets.

Le nombre d'arrangements de (**p**) objets pris parmi (**n**) est noté : $\boldsymbol{A_n^p}$

a) Arrangement avec répétition :

Lorsqu'un objet peut être observé **plusieurs fois** dans un arrangement, le nombre **d'arrangement avec répétition** de (**p**) objets pris parmi (**n**), est alors : $\boldsymbol{A_n^p = n^p}$ (1≤p≤n)

b) Arrangement sans répétition :

Lorsque chaque objet ne peut être observé **qu'une seule fois** dans un arrangement, le nombre **d'arrangements sans répétition** de (**p**) objets pris parmi (**n**) est alors : $\boldsymbol{A_n^p = \frac{n!}{(n-p)!}}$

$1 \times 2 \times 3 \times \ldots \times p \times (p+1) \times \ldots \times (n-1) \times n = n!$

4.1.2 Permutations :

a) Permutations sans répétitions :

Etant donné un ensemble **E** de (**n**) objets, on appelle **permutations** de (**n**) objets distincts **toutes suites ordonnées** de (**n**) objets ou tout **arrangement n à n** de ces objets. Le nombre de permutations de (**n**) objets est noté : $\boldsymbol{P_n = A_n^n = n!}$

b) Permutations avec répétitions :

Dans le cas où il existerait plusieurs répétitions (**k**) d'un même objet parmi les (**n**) objets, le nombre de permutations possibles des (**n**) objets doit être rapporté aux nombres de permutations des (**k**) objets identiques.

Le nombre de permutations de n objets est alors :$\boldsymbol{P_n = \frac{n!}{k!}}$

4.2.3 Combinaisons (désordre) :

Si l'on reprend la définition des arrangements où on parle d'une seule suite ordonnée, c'est-à-dire une seule combinaison. Dans les combinaisons, on ne parle plus de suite ni de série puisque la notion d'ordre des objets n'est plus prise en compte. On parle alors de tirages avec ou sans remise.

a) Combinaisons sans remise :

Etant donné un ensemble **E** de (**n**) objets, on appelle **combinaisons** de (**p**) objets tout ensemble de (**p**) objets pris parmi les (**n**) objets sans remise.
Le nombre de combinaisons de (**p**) objets pris parmi (**n**) est noté **:** $\boldsymbol{C_n^p = \frac{n!}{P!(n-p)!}}$

b) Combinaisons avec remise :

Le nombre de combinaisons de (**p**) objet parmi (**n**) avec remise est : **:** $\boldsymbol{C_{n+p-1}^p = \frac{(n+p-1)!}{P!(n-1)!}}$

4.3 Notions de bases :

La théorie des probabilités fournit des modèles mathématiques permettant l'étude d'expériences dont le résultat ne peut être prévu avec une totale certitude. Autrement dit, calculer une probabilité revient donc à quantifier la possibilité qu'un évènement se produise lors d'une expérience qui ne découle que du hasard. Voici quelques notions de bases en probabilités :

4.3.3 Expérience aléatoire

Une expérience aléatoire est un processus dans lequel intervient le hasard et qui est susceptible de produire différents résultats

Exemples des expériences aléatoires :

1- Lancer deux dés et observer le total
2- Tirer le numéro gagnant d'une loterie
3- Jeter une pièce de monnaie deux fois et noter le coté qui apparait

4.3.4 Espace échantillonnai

L'espace échantillonnai d'une expérience aléatoire est l'ensemble de tous les résultats possibles de cette expérience, noté Q

Exemples

Les espaces échantillonnaux associés aux expériences aléatoires présentées dans l'exemple précédent sont respectivement :

1. Ω = {2,3,4,5,6,7,8,9,10, 11,12}
2. Ω = {1000,1001, ...,9999}
3. Ω = {ff, fp, pf, pp}

4.3.5 Evénement

Un événement relié à une expérience aléatoire est un sous-ensemble de l'espace échantillonnai Q. On note habituellement les événements par *A, B, C,...*

Exemple

Soit l'expérience consistant à jeter une pièce de monnaie deux fois et de noter le côté qui apparait. Ainsi, l'espace échantillonnai est

B = {ff,fp,pf,pp}

Voici quelques exemples d'événements :

A = " obtenir face au premier lancer " = *{fp.ff} ;*

B = " obtenir face au deuxième lancer " = *{pf. ff} ;*

C = " obtenir le même côté lors des deux lancers " = *{pp, ff} ;*

D = " obtenir des côtés différents lors des deux lancers ” = *{pf, fp] ;*

À l'aide des opérations sur les ensembles, nous pouvons, à partir d'un ou de plusieurs événements, en former de nouveaux.

Si 4 et *B* sont deux événements, alors :

1. $\bar{A}$ est l'événement qui se réalise si l'événement *A ne se* réalisé pas. On dit que $\bar{A}$ est l'événement complémentaire de l’événement *A*.
2. *A*∩ *B* est l'événement pour lequel les deux événements se réalisent (Si A ∩B = 0, on dit que A et B sont des événements mutuellement exclusifs)
3. AUB est l'événement pour lequel au moins un des événements A ou B se réalise.
4. A \ *B* est l'événement pour lequel A est réalisé mais non B.

4.3 Calcul des probabilités

La probabilité P(A) de réaliser un événement A est égale au rapport du nombre n de cas favorables à la réalisation de cet événement au nombre N de cas possibles :

$$P(A) = \frac{Nombre\ de\ cas\ favorables}{Nombre\ de\ cas\ possibles} = \frac{n}{N} = \frac{A}{\Omega}$$

Propriétés des probabilités

Soient *A* et *B,* des événements quelconques. Alors, les propriétés suivantes doivent être satisfaites :

1. $0 < P(A) < 1$
2. $P(\Omega) = 1$
3. $P(\bar{A}) = 1 - P(A)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(A \setminus B) = P(4) - P(A \cap B)$

Nous pouvons déduire en combinant les propriétés 2 et 3 que la probabilité d'obtenir l'ensemble vide est nulle : $P(\emptyset) = 1 - P(\Omega) = 0$

4.5 Probabilités conditionnelles :

Il est question de probabilités conditionnelles dès que nous sommes intéressés à la probabilité qu'un événement A se produise, *sachant qu'un autre événement B est réalisé.* Nous noterons par $P(A \mid B)$ cette probabilité. En quelque sorte, ce type de probabilité nous oblige *à* considérer *B* (plutôt que Ω) comme étant l'espace échantillonnal duquel nous étudions les chances de réalisation de *A.*

Pour tout A et B, éléments de Ω

$$P(A \setminus B) = \frac{P(A \cap B)}{P(B)}$$

4.6 Evènement indépendant :

Soient deux événements 4 et B. Nous disons de 4 et B qu'ils sont des *événements indépendants* si et seulement si

$$\mathbf{P(A\backslash}\boldsymbol{B)=P(A)}$$

Cette définition peut être interprétée de la façon suivante : si la réalisation de l'événement B ne modifie pas la probabilité que l'événement A se produise, alors A et B sont indépendants.

4.7 Théorème de Bayes

Le théorème de Bayes est une conséquence immédiate des probabilités conditionnelles et des probabilités totales. Probabilités conditionnelles .

Exemple : Dans une bibliothèque comportant 100 ouvrages, il y en a 40 qui sont écrits en anglais dont 8 portent sur la biologie. Considérons les événements suivants :

A = “le livre est écrit en anglais” ; P(A) = 40 /100

B = “le livre porte sur la biologie” ;

A ∩ B = “le livre est écrit en anglais et porte sur la biologie” ; P(A ∩ B) = 8 / 100

Probabilité conditionnelle B | A = “le livre porte sur la biologie sachant qu’il est écrit en anglais” ; P(B | A) = 8/ 40 Il s’agit de la fréquence des livres de biologie parmi les livres en langue anglaise.

On a les relations $P(B \mid A) = \frac{\frac{8}{100}}{\frac{40}{100}} = \frac{8}{40} = \frac{P(A \cap B)}{P(A)}$ Retenons $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$ ou encore

$P(A \cap B) = P(A) \cdot P(B \mid A)$

Formule de Bayes : Considérons une partition A1, A2, ..., An de l’ensemble des événements E. Alors :

$$P(B) = P(A_1) \cdot P(B \mid A_1) + P(A_2) \cdot P(B \mid A_2) + \ldots + P(A_n) \cdot P(B \mid A_n)$$

$$P(A_1 \mid B) = \frac{P(A_1) \cdot P(B \mid A_1)}{P(B)}$$

$$P(A_2 \mid B) = \frac{P(A_2) \cdot P(B \mid A_2)}{P(B)}$$

$$\ldots$$

$$P(A_n \mid B) = \frac{P(A_n) \cdot P(B \mid A_n)}{P(B)}$$

5. CHAPITRE 4 : VARIABLES ALÉATOIRES ET LES PRINCIPALES LOIS DE PROBABILITÉ

5.1 Notion de variables aléatoires :

Une variable aléatoire X est une variable associée à une expérience ou à un groupe d'expériences aléatoires et servant à caractériser le résultat de cette expérience ou de ce groupe d'expériences. On distingue les variables aléatoires discontinues ou discrètes et les variables aléatoires continues.

5.2 Variable aléatoire discontinue :

Une variable aléatoire est discrète si elle varie de façon discontinue, la variable ne peut prendre que des valeurs entières.

Soit x une variable pouvant prendre l'ensemble des valeurs :

X1,X2,……..Xn avec les probabilités P1,P2,……,Pn respectivement, telles que :

P1+P2+……+Pn=1 $\sum Pi = 1$

5.2.1 Loi de probabilité :

On dit qu'on a défini une loi probabilité (fonction de distribution) d'une variable aléatoire discontinue si on arrive à déterminer toutes les valeurs que peut prendre la variable Xi et toutes les probabilités correspondantes Pi**.**

Esperance mathematique : l'ésperance mathematique d'une variable aléatoire discontinue, noté E(x)

Telque **E(x)=∑PiXi**

La variance : la variance d'une variable aléatoire discontinue, noté V(x)

Telque **V(x)= ∑Pi(Xi-E(xi))^2**

La distribution cumulée des probabilités est appelée fonction de répartition :

$$F(x) = p(X \leq x) = \sum^{x} p(x)$$

$$0 \leq F(x) \leq 1$$

Exemple :

Soit X la variable aléatoire qui caractérise le résultat de l'expérience aléatoire "jet d'un dé homogène". X est une variable aléatoire discrète, elle peut prendre les valeurs entières 1, 2, 3, 4, 5, et 6 avec la probabilité constante 1/6.

x	p(x)	F(x)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6
Total	1	

Remarque : Le but des lois théoriques est la description des phénomènes statistiques dont le but de calculer la probabilité de certains événements et donc d'avoir une certaine représentation de l'avenir.

Nous étudierons au cours de ce chapitre les lois de probabilités les plus courantes qui vont nous permettre la description d'un phénomène aléatoire déterminé. Nous présenterons ainsi la loi de Bernoulli, la loi binomiale, et la loi de poisson.

5.2.1.1 Loi de Bernoulli :

La loi de Bernoulli intervient dans le cas d'une seule expérience aléatoire à laquelle on associe un événement aléatoire quelconque.

La réalisation de l'événement au cours de cette expérience est appelée succès et la probabilité de réalisation est dite probabilité de succès, désignée par **p**. Par contre la non-réalisation de l'événement est appelée échec et la probabilité de non-réalisation est dite probabilité d'échec, désignée par **q**. **q = 1 - p**

La variable aléatoire X qui caractérise le nombre de succès au cours d'une seule expérience aléatoire est appelée variable de Bernoulli, elle prend les valeurs entières 0 et 1 avec les probabilités respectivement q et p.

Les caractéristiques d'une variable Bernoulli sont :

- **Espérance mathématique**

$$E(X) = \sum xp(x) = 0 \times q + 1 \times p = p$$

- **Variance**

$$E(X^2) = \sum x^2 p(x) = 0^2 \times q + 1^2 \times p = p$$

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p\,(1 - p) = pq$$

5.2.1.2 Loi Binomiale :

La loi binomiale intervient dans le cas de plusieurs expériences aléatoires identiques et indépendantes auxquelles on associe un événement aléatoire quelconque.

La réalisation de l'événement au cours de chacune des expériences est appelée succès et la probabilité de réalisation est dite probabilité de succès, désignée par p. Par contre la non réalisation de l'événement est appelée échec et la probabilité de non-réalisation est dite probabilité d'échec, désignée par q. $q = 1 - p$

Les probabilités p et q restent constantes au cours d'une suite d'expériences aléatoires.

La variable aléatoire X qui caractérise le nombre de succès au cours de n expériences aléatoires indépendantes est appelée variable binomiale, elle prend les valeurs entières de 0 à n.

La probabilité d'obtenir x succès et donc (n-x) échecs au cours de n expériences aléatoires indépendantes est, pour x = 0, 1, ..., n :

$$p(x) = C_n^x p^x q^{n-x}$$

La loi binomiale dépend de deux paramètres :

- n = nombre d'expériences aléatoires indépendantes ;
- p = probabilité de succès au cours de chacune des n expériences aléatoires, p doit rester constante.

Une variable aléatoire X qui suit une loi binomiale de paramètres n et p, est désignée par :

$X = B(n , p)$

$$E(x)=np$$

$$V(x)=npq$$

5.2.1.3 Loi de poisson :

La loi de poisson intervient pour des phénomènes statistiques dont le nombre de réalisation varie de 0 à l'infini et dont la fréquence moyenne de réalisation est connue.

Exemple :

Nombre d'appels reçus par un standard téléphonique.

Nombre d'accidents de la circulation.

Nombre de visiteur d'un centre commercial.

La variable aléatoire X qui caractérise le nombre de réalisations de ce phénomène est appelée variable de poisson, elle prend les valeurs entières 0,1, 2, …etc.

La probabilité d'obtenir x réalisations est, pour x = 0, 1, 2

$$p(x) = \frac{e^{-m} \times m^{x}}{x!}$$

La loi binomiale dépend d'un seul paramètre :

- m = fréquence moyenne du phénomène étudié.

Une variable aléatoire X qui suit une loi de poisson de paramètre m est désignée par :

$$X = P(m)$$

5.3 Variable aléatoire continue :

Lorsque l'on s'intéresse à la durée d'une communication téléphonique, à la durée de vie d'un composant électronique ou à la température de l'eau d'un lac, la variable aléatoire X associée au temps ou à la température, peut prendre une infinité de valeurs dans un intervalle donné. On dit alors que cette variable X est continue (qui s'oppose à discrète comme c'est le cas par exemple dans la loi binomiale)

On ne peut plus parler de probabilité d'événements car les événements élémentaires sont en nombre infini. La probabilité d'une valeur isolée de X est alors nulle. On contourne cette difficulté en associant à la variable X un intervalle de R et en définissant une densité de probabilité.

Elle est caractérisée par deux paramètres qui sont la moyenne et l 'écart type et l 'écart type.

5.3.1 Densité de probabilité et espérance mathématique

Définition : On appelle densité de probabilité d'une variable aléatoire continue X, toute fonction f continue et positive sur un intervalle I ([a; b], [a ;+oo[ou R) telle que :

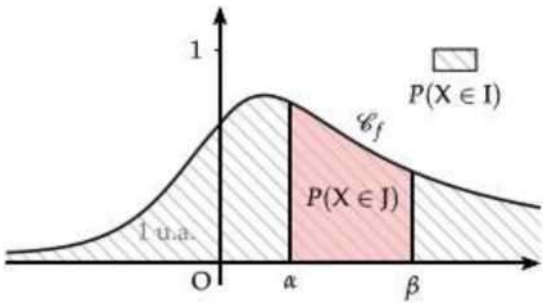
P(X=I)=$\int_{(I)} \boldsymbol{f(t)dt = 1}$

Pour tout intervalle J=[α,β] inclus dans I, on a **:** $\boldsymbol{\int_{\alpha}^{\beta} f(t)dt}$

D'autre part la fonction F définie par : F(x)=P(X<x) est appelée la fonction de répartition de la variable X

$\boldsymbol{\int_{a}^{x} f(t)dt}$ **ou** $\boldsymbol{\lim_{a<-\infty} \int_{a}^{x} f(t)dt}$

- **Remarque :** Comme la fonction f est continue et positive, la probabilité P(X ∈ I) correspond à l'aire sous la courbe φ_f Elle vaut alors 1 .
- La probabilité P(X ∈ J), avec J = [α; β], correspond à l'aire du domaine délimité par φ_f, l'axe des abscisse et les droites d'équation x =α et y =β



Définition :

L'espérance mathématique d'une variable aléatoire continue X, de densité f sur I, est :

E(x)= $\int_{(I)} tf(t)dt$

Lien entre le discret et le continu

Discret	Continu
Univers Ω	Intervalle I ou ℝ
Événement E sous-ensemble de Ω	Événement J sous-intervalle de I
Probabilités p_i des événements élémentaires $\sum p_i = 1$	Densité de probabilité $\int_{(I)} f(t)\,dt = 1$
Espérance de la variable aléatoire X $E(X) = \sum p_i x_i$	Espérance de la variable aléatoire X $E(X) = \int_{(I)} t\,f(t)\,dt$
Équiprobabilité $P(E) = \frac{\text{nbre de cas favorables}}{\text{nbre de cas possibles}}$	Loi uniforme $P(X \in J) = \frac{\text{longueur de J}}{\text{longueur de I}}$

5.3.2 Loi normale

Définition :

La loi normale est la loi continue la plus importante et la plus utilisée dans le calcul de probabilité. Elle est aussi appelée loi de LAPLACE GAUSS[1].

On appelle variable normale toute variable aléatoire continue X définie dans l'intervalle]-∞ ;+∞[par la fonction de densité de probabilité suivante :

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$$

m et Σ sont des paramètres quelconques qui représentent respectivement la moyenne et l'écart type de la variable.

On peut vérifier que :

$$\int_{-\infty}^{+\infty} f(x)\mathrm{dx} = 1$$

La loi normale dépend de deux paramètres m et σ. Une variable aléatoire X qui suit une loi normale de paramètres met G est désignée par :

$$X = N(m , \sigma)$$

5.3.2.1 Loi normale réduite

On appelle variable normale réduite toute variable aléatoire normale Z de paramètres m = 0 et σ = 1. Z = N(0 , 1)

Une variable normale réduite est définie par la fonction de densité de probabilité suivante :

$$F(z) = \frac{1}{\sqrt{2\Pi}} e^{\frac{z^2}{2}}$$

Toute variable normale X de paramètres m et σ peut être transformée en une variable normale réduite par le changement de variable suivant :

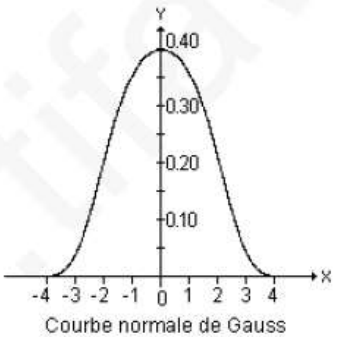
$$Z=\frac{x-m}{\sigma}$$

Forme de la loi normale

La représentation graphique de la fonction de densité de probabilité d'une variable normale est une courbe en forme de cloche symétrique par rapport à la moyenne m et caractérisée par l'existence d'un maximum

$$x=0 \text{ et } f(x)=\frac{1}{\partial\sqrt{2\pi}}$$

En particulier la loi normale réduite est symétrique par rapport à l'axe des abscisses et caractérisée par l'existence d'un maximum en **z=0** et $f(z)=\frac{1}{\sqrt{2\pi}}= 0\ ;40$



Courbe normale de Gauss

La fonction de répartition correspond à l'aire comprise entre cette courbe et l'axe des abscisses.

Détermination pratique des probabilités

Pour le calcul de probabilités sans utiliser la fonction de densité, des tables de la loi normale réduite ont été élaborées. On distingue deux tables de la loi normale réduite, relatives l'une à la fonction de densité de probabilité et l'autre à la fonction de répartition. En raison de la symétrie de la distribution, ces tables sont limitées aux valeurs positives de z.

Par le changement de variable Z toutes les variables normales se ramènent à la loi normale réduite.

Table de la fonction de répartition

Cette table donne les valeurs de la fonction de répartition Π(z) pour des valeurs positives z d'une variable normale réduite. En raison de la symétrie de f(z). on peut déduire les valeurs n(z) pour les valeurs négatives de z :

$$\Pi(-z) = p(Z \leq -z) = p(Z > z) = 1 - p(Z \leq z) = 1 - \Pi(z)$$

$$\Pi(-z) = 1 - \Pi(z)$$

Pour une variable normale quelconque X de paramètre m et σ :

$$F(x)=p(X \leq x)=p\left(\frac{X-m}{\sigma} \leq \frac{x-m}{\sigma}\right)=p(Z<z)=\Pi(z)$$

$$F(X) = \Pi(z)$$

Pour lire une valeur Π(z) dans la table, il suffit de lire l'intersection entre la ligne correspondante à la valeur de z et la colonne correspondante au deuxième chiffre après la virgule de z

TABLE DE LA FONCTION DE REPARTITION DE LA LOI NORMALE REDUITE

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	05359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	05753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	06141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	06517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	06879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	07224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	07549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	07852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	08133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	08389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	08621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	08830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	090147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99897	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

Exemple :

La valeur de $\Pi(1,36)$ correspond à l'intersection entre la ligne correspondante à 1,3 et la colonne correspondante à 0,06, on peut lire la valeur 0,91309.

$$\Pi(-2,24) = 1 - \Pi(2,24) = 1 - 0,98745 = 0,01255$$

Exemple :

Pour qu'une pièce fabriquée par une machine soit utilisable, sa longueur doit être comprise entre 14,7 et 15,3 cm. sinon elle est rejetée. Sachant que la longueur de cette pièce est une variable normale de paramètres 15 cm et 0,2 cm, quelle proportion de pièces peuvent être rejetées.

Si on désigne par h variable X la longueur des pièces. X suit une loi normale :

$$X = N(15; 0,2)$$

La probabilité de rejet d'une pièce est :

p(rejet) = 1 -p(accepter)

p(accepter) = p(14,7≤X≤ 15,3) = p(X≤ 15,3)-p(X≤ 14.7)

$$p(\text{accepter}) = p\left(\frac{X-15}{0.2} \leq \frac{15,3-15}{0.2}\right) - p\left(\frac{X-15}{0.2} \leq \frac{14,7-15}{0.2}\right)$$

p(accepter) = p(Z < 1.50) - p(Z < -1,50)

p(accepter) = II (1,50) - II(-l,50)

p(accepter) = II (1,50) - (1 - II (1.50)) = 2 II (l,50) - 1

p(accepter) = 2x 0.93319- 1 = 0.86638

Chaque pièce a une probabilité de 0.13362 d'être rejetée ou il y a un risque de rejet de 13% des pièces fabriquées

5.3.3 Loi Khi-deux :

Définition :

On appelle variable Khi deux de Pearson. la variable x^2 qui varie entre 0 et +∞ et définie par la fonction de densité de probabilité :

$$f(x)=cx^{\frac{k}{2}-1}e^{\frac{-x}{2}}$$

Le paramètre k est une constante entière positive appelée nombre de degrés de liberté, on dit variable Khi carré à k degré de liberté, désignée par x^2à k dl

C'est une constante telle que : $\int_0^{+\infty} f(x)dx = 1$

La variable Khi deux de Pearson correspond aussi à la somme des carrés de k variables normales réduites indépendantes.

Table de khi-deux :

k / p	0,0005	0,001	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,4
1	$0,0^6393$	$0,0^5157$	$0,0^4393$	$0,0^3157$	$0,0^3982$	$0,0^2393$	0,0158	0,0642	0,148	0,275
2	$0,0^2100$	$0,0^2200$	0,0100	0,0201	0,0506	0,103	0,211	0,446	0,713	1,02
3	0,0153	0,0243	0,0717	0,115	0,216	0,352	0,584	1,00	1,42	1,87
4	0,0639	0,0908	0,207	0,297	0,484	0,711	1,06	1,65	2,19	2,75
5	0,158	0,210	0,412	0,554	0,831	1,15	1,61	2,34	3,00	3,66
6	0,299	0,381	0,676	0,872	1,24	1,64	2,20	3,07	3,83	4,57
7	0,485	0,598	0,989	1,24	1,69	2,17	2,83	3,82	4,67	5,49
8	0,710	0,857	1,34	1,65	2,18	2,73	3,49	4,59	5,53	6,42
9	0,972	1,15	1,73	2,09	2,70	3,33	4,17	5,38	6,39	7,36
10	1,26	1,48	2,16	2,56	3,25	3,94	4,87	6,18	7,27	8,30
11	1,59	1,83	2,60	3,05	3,82	4,57	5,58	6,99	8,15	9,24
12	1,93	2,21	3,07	3,57	4,40	5,23	6,30	7,81	9,03	10,2
13	2,31	2,62	3,57	4,11	5,01	5,89	7,04	8,63	9,93	11,1
14	2,70	3,04	4,07	4,66	5,63	6,57	7,79	9,47	10,8	12,1
15	3,11	3,48	4,60	5,23	6,26	7,26	8,55	10,3	11,7	13,0
16	3,54	3,94	5,14	5,81	6,91	7,96	9,31	11,2	12,6	14,0
17	3,98	4,42	5,70	6,41	7,56	8,67	10,1	12,0	13,5	14,9
18	4,44	4,90	6,26	7,01	8,23	9,39	10,9	12,9	14,4	15,9
19	4,91	5,41	6,84	7,63	8,91	10,1	11,7	13,7	15,4	16,9
20	5,40	5,92	7,43	8,26	9,59	10,9	12,4	14,6	16,3	17,8
21	5,90	6,45	8,03	8,90	10,3	11,6	13,2	15,4	17,2	18,8
22	6,40	6,98	8,64	9,54	11,0	12,3	14,0	16,3	18,1	19,7
23	6,92	7,53	9,26	10,2	11,7	13,1	14,8	17,2	19,0	20,7
24	7,45	8,08	9,89	10,9	12,4	13,8	15,7	18,1	19,9	21,7
25	7,99	8,65	10,5	11,5	13,1	14,6	16,5	18,9	20,9	22,6
26	8,54	9,22	11,2	12,2	13,8	15,4	17,3	19,8	21,8	23,6
27	9,09	9,80	11,8	12,9	14,6	16,2	18,1	20,7	22,7	24,5
28	9,66	10,4	12,5	13,6	15,3	16,9	18,9	21,6	23,6	25,5
29	10,2	11,0	13,1	14,3	16,0	17,7	19,8	22,5	24,6	26,5
30	10,8	11,6	13,8	15,0	16,8	18,5	20,6	23,4	25,5	27,4

k / p	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,455	0,708	1,07	1,64	2,71	3,84	5,02	6,63	7,88	10,8	12,1
2	1,39	1,83	2,41	3,22	4,61	5,99	7,38	9,21	10,6	13,8	15,2
3	2,37	2,95	3,67	4,64	6,25	7,81	9,35	11,3	12,8	16,3	17,7
4	3,36	4,04	4,88	5,99	7,78	9,49	11,1	13,3	14,9	18,5	20,0
5	4,35	5,13	6,06	7,29	9,24	11,1	12,8	15,1	16,7	20,5	22,1
6	5,35	6,21	7,23	8,56	10,6	12,6	14,4	16,8	18,5	22,5	24,1
7	6,35	7,28	8,38	9,80	12,0	14,1	16,0	18,5	20,3	24,3	26,0
8	7,34	8,35	9,52	11,0	13,4	15,5	17,5	20,1	22,0	26,1	27,9
9	8,34	9,41	10,7	12,2	14,7	16,9	19,0	21,7	23,6	27,9	29,7
10	9,34	10,5	11,8	13,4	16,0	18,3	20,5	23,2	25,2	29,6	31,4
11	10,3	11,5	12,9	14,6	17,3	19,7	21,9	24,7	26,8	31,3	33,1
12	11,3	12,6	14,0	15,8	18,5	21,0	23,3	26,2	28,3	32,9	34,8
13	12,3	13,6	15,1	17,0	19,8	22,4	24,7	27,7	29,8	34,5	36,5
14	13,3	14,7	16,2	18,2	21,1	23,7	26,1	29,1	31,3	36,1	38,1
15	14,3	15,7	17,3	19,3	22,3	25,0	27,5	30,6	32,8	37,7	39,7
16	15,3	16,8	18,4	20,5	23,5	26,3	28,8	32,0	34,3	39,3	41,3
17	16,3	17,8	19,5	21,6	24,8	27,6	30,2	33,4	35,7	40,8	42,9
18	17,3	18,9	20,6	22,8	26,0	28,9	31,5	34,8	37,2	42,3	44,4
19	18,3	19,9	21,7	23,9	27,2	30,1	32,9	36,2	38,6	43,8	46,0
20	19,3	21,0	22,8	25,0	28,4	31,4	34,2	37,6	40,0	45,3	47,5
21	20,3	22,0	23,9	26,2	29,6	32,7	35,5	38,9	41,4	46,8	49,0
22	21,3	23,0	24,9	27,3	30,8	33,9	36,8	40,3	42,8	48,3	50,5
23	22,3	24,1	26,0	28,4	32,0	35,2	38,1	41,6	44,2	49,7	52,0
24	23,3	25,1	27,1	29,6	33,2	36,4	39,4	43,0	45,6	51,2	53,5
25	24,3	26,1	28,2	30,7	34,4	37,7	40,6	44,3	46,9	52,6	54,9
26	25,3	27,2	29,2	31,8	35,6	38,9	41,9	45,6	48,3	54,1	56,4
27	26,3	28,2	30,3	32,9	36,7	40,1	43,2	47,0	49,6	55,5	57,9
28	27,3	29,2	31,4	34,0	37,9	41,3	44,5	48,3	51,0	56,9	59,3
29	28,3	30,3	32,5	35,1	39,1	42,6	45,7	49,6	52,3	58,3	60,7
30	29,3	31,3	33,5	36,3	40,3	43,8	47,0	50,9	53,7	59,7	62,2

Pour lire une valeur x^2à k dl dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de la probabilité cumulée F(x^2à k dl) et la ligne correspondante aux degrés de liberté k.

Exemple :

La valeur de x^2â 10 dl pour une probabilité de 0,95 correspond à l'intersection entre la colonne correspondante à 0.95 et la ligne correspondante à 10. on peut lire la valeur 18,3.

$X^2 0.95$ à 10dl = 18.3

$X^2 0{,}05$ à 20dl = 10,9

5.3.4 Loi de Student :

On appelle variable t de Student, la variable t qui varie entre -∞ et +∞ et définie par la fonction de densité de probabilité :

$$\boldsymbol{f(t) = c(1 + \frac{t^2}{k})^{\frac{k+1}{2}}}$$

Le paramètre k est une constante entière positive appelée nombre de degrés de liberté, on dit variable t à k degré de liberté, désignée par t à k dl.

C'est une constante telle que : $\int_0^{+\infty} \boldsymbol{f(x)dx = 1}$

La variable t de Student correspond aussi au quotient d'une variable normale réduite par la racine carrée d'une variable x^2 à k dl indépendante de la première variable.

Soient Z une variable normale réduite et x^2 à k dl une variable Khi carré à k degrés de liberté, indépendantes. On peut démontrer :

$$\boldsymbol{T\grave{a}kdl = \frac{Z}{\sqrt{\frac{x^2\ \grave{a}kdl}{k}}}}$$

Table de Student :

La table de la loi t de Student dépend du paramètre k, elle donne les valeurs de t à k dl pour les valeurs de la fonction de répartition F(t à k dl).

k / p	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Pour lire une valeur t_{ikdl} dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de la probabilité cumulée $F(t_{i\,k\,dl})$ et la ligne correspondante aux degrés de liberté k.

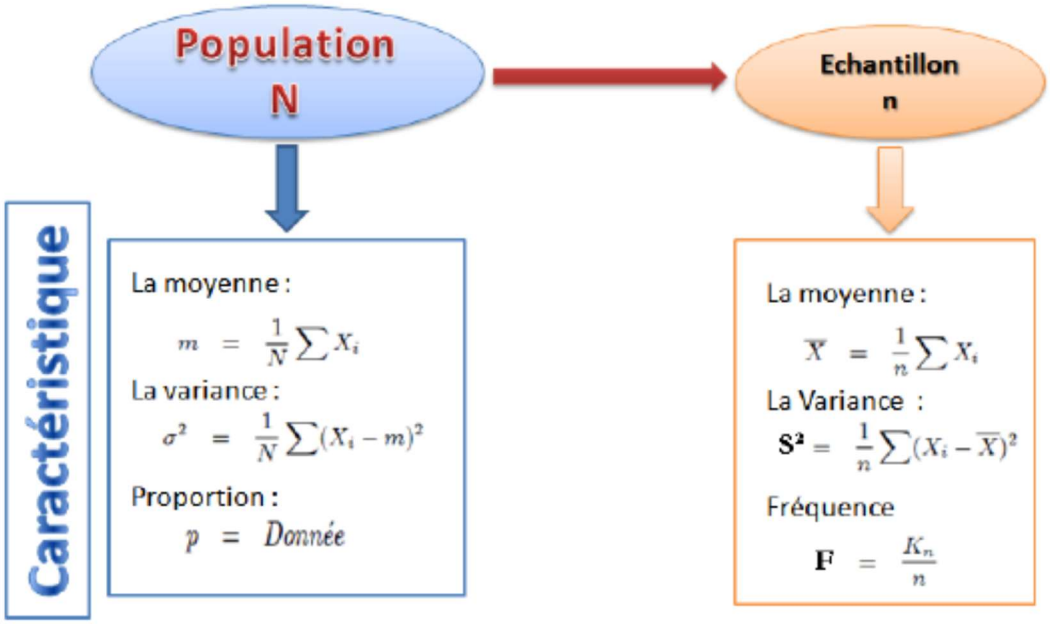
6. CHAPITRE 5 : LES DISTRIBUTIONS D'ÉCHANTILLONNAGE

6.1 Introduction

L'objectif de cette partie est de répondre à la problématique suivante : *comment, à partir d'informations (couple moyenne-écart-type ou proportion) connues sur une population, peut-on prévoir celles d'un échantillon ?*

Le verse versa sera dans la partie d'estimation, c'est elle qui va répondre à la question : comment à partir d'une étude sur l'échantillon on peut avoir une idée sur la population ?

Puisque on a trois caractéristiques de population, alors on aura 3 types de distributions d'échantillonnage : Moyenne, Variance et Proportion.



6.2 Type des échantillons :

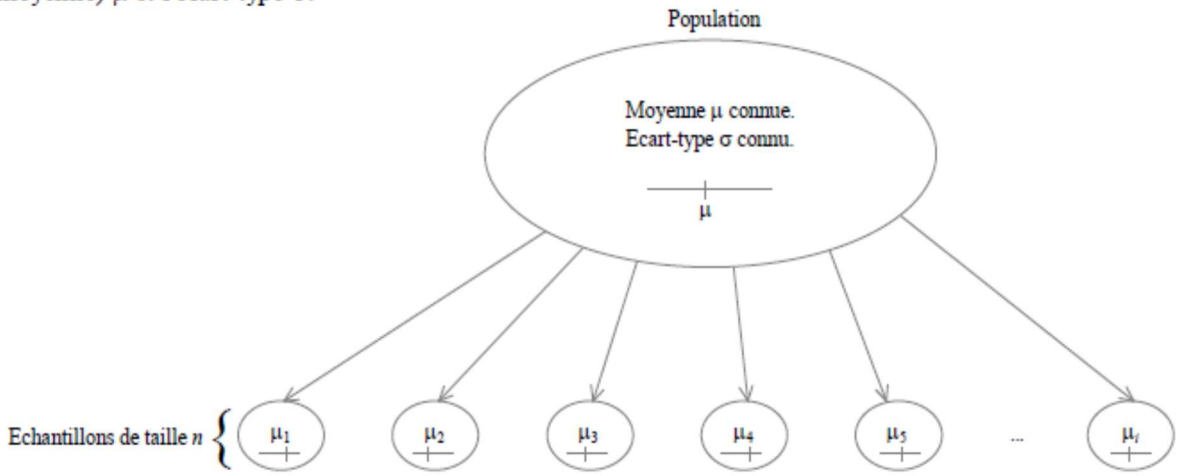
Dans notre cours, nous allons travailler sur l'échantillonnage aléatoire simple, avec deux cas :

1. Non Exhaustif : Avec Remise car la taille de la population est grande.

2. Exhaustif : Sans Remise car la taille de population est finie.

6.3 Distributions d'échantillonnage de la Moyenne $\overline{X}$:

On dispose d'une population sur laquelle est définie une variable aléatoire X dont on connaît l'espérance (ou la moyenne) μ et l'écart-type σ.



On s'intéresse aux échantillons de taille n. Auront-ils tous la même moyenne ? Non, certains peuvent être constitués d'éléments atypiques et avoir une moyenne très différente de celle de la population (surtout si l'échantillon est de petite taille).

Notons $\overline{X}$ la variable aléatoire qui, à chaque échantillon de taille n, associe sa moyenne ($\overline{X}$ s'appelle encore la *distribution des moyennes des échantillons*). Que peut-on dire de cette variable aléatoire $\overline{X}$?

Soit une population, et soit X V.A telque $E(X) = m$ et $var(X) = \sigma^2$, et un échantillon de taille n, Et on sait que la moyenne de l'échantillon est

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Alors puisque $E(X) = m$ et $var(X) = \sigma^2$, on a :

1. **La moyenne de $\overline{X}$:**

$$E(\overline{X}) = m.$$

2. **La variance de $\overline{X}$:**

Pour la variance on a deux cas : (ça dépend des types des échantillons)

(a) **Non Exaustif : Avec Remise**

$$\begin{aligned} Var(\overline{X}) &= \frac{\sigma^2}{n} \\ \sigma_{\overline{X}} &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

(b) **Exaustif : Sans Remise** (Taille N sera donnée)

$$\begin{aligned} Var(\overline{X}) &= \frac{\sigma^2}{n} \times \frac{N-n}{N-1} \\ \sigma_{\overline{X}} &= \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \end{aligned}$$

3. **La loi de $\overline{X}$:**

Pour la loi de $\overline{X}$ on a deux cas (ça dépend de σ)

(a) **Si σ (ou σ^2) connue (Donnée) :**

Non Exaustif : Avec Remise

$$\begin{aligned} \overline{X} &\rightsquigarrow N(m, \frac{\sigma}{\sqrt{n}}) \\ \text{Donc } T &= \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1) \end{aligned}$$

Exaustif : Sans Remise

$$\begin{aligned} \overline{X} &\rightsquigarrow N(m, \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}) \\ \text{Donc } T &= \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \rightsquigarrow N(0,1) \end{aligned}$$

(b) **Si σ (ou σ^2) inconnue (Non Donnée) :**

$$\overline{X} \rightsquigarrow N(m, \frac{s}{\sqrt{n}})$$

(Puisque σ est inconnue on la remplace par s)

$$\text{Donc } T = \frac{\overline{X} - m}{\frac{s}{\sqrt{n}}} \rightsquigarrow t(n-1) \text{ [Student de n – 1 degret de Libérté]}$$

Démonstration :

Notons $E = \{x_1 ; x_2 ; ... ; x_n\}$ un échantillon de n éléments prélevés au hasard dans la population.

Pour tout i compris entre 1 et n, notons X_i la variable aléatoire correspondant à la valeur du i-ème élément x_i de l'échantillon. Nous savons, par hypothèse, que :

$$E(X_i) = \mu \text{ et } \sigma(X_i) = \sigma$$

La moyenne $\overline{X}$ des n valeurs de l'échantillon est :

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

D'après les propriétés de la loi normale, nous savons qu'une combinaison linéraire de variables aléatoire qui suivent la loi normale est encore une variable aléatoire qui suit la loi normale. Comme chaque variable aléatoire X_i suit ici la loi normale $N(\mu, \sigma)$, la variable aléatoire moyenne $\overline{X}$ suit donc également une loi normale. Calculons ses paramètres.

D'après la propriété de linéarité de l'espérance :

$$E\left(\overline{X}\right) = \frac{E(X_1) + E(X_2) + ... + E(X_n)}{n} = \frac{n\mu}{n} = \mu$$

D'après les propriétés de la variance :

$$V\left(\overline{X}\right) = \frac{V(X_1) + V(X_2) + ... + V(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

D'où :

$$\sigma\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}$$

Exemple :

Les statistiques des notes obtenues en mathématiques au BAC STI en France pour l'année 2006 sont :

Moyenne nationale : $\mu = 10{,}44$

Écart-type : $\sigma = 1{,}46$

Une classe de BTS comporte 35 élèves en 2006/2007 issus d'un BAC STI en 2006.

Calculer la probabilité que la moyenne de cette classe soit supérieure à 10.

Ici, nous ne connaissons pas la loi sur la population, mais l'effectif n de l'échantillon est supérieur à 30.

Nous allons donc pouvoir utiliser le T.C.L. 2.

Notons $\overline{X}$ la variable aléatoire qui, à tout échantillon de taille $n = 35$, fait correspondre sa moyenne.

Alors : $$\overline{X} \rightsquigarrow N\left(\mu ; \frac{\sigma}{\sqrt{n}}\right) = N\left(10,44 ; \frac{1,46}{\sqrt{35}}\right)$$

Posons $T = \dfrac{\overline{X} - 10,44}{\dfrac{1,46}{\sqrt{35}}}$ ainsi $T \rightsquigarrow N(0 ; 1)$.

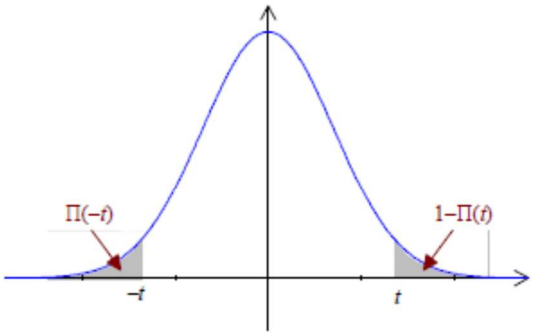
Nous obtenons alors par centrage et réduction :

$$P(\overline{X} \geqslant 10) = P\left(\frac{\overline{X} - 10,44}{\dfrac{1,46}{\sqrt{35}}} \geqslant \frac{10 - 10,44}{\dfrac{1,46}{\sqrt{35}}}\right)$$

$$= P(T \geqslant -1,78)$$

$$= P(T \leqslant 1,78)$$

$$= \Pi(1,78)$$



Remarque : $P(T \geqslant t) = P(T \leqslant -t)$

En effet :

$P(T \geqslant t) = 1 - P(T \leqslant t) = 1 - \Pi(t) = \Pi(-t) = P(T \leqslant -t)$

Et par lecture directe de la table de la loi normale centrée-réduite :

$$\Pi(1,78) = 0,9625$$

Conclusion : il y a environ 96% de chance que, dans cette classe de BTS, la moyenne des notes au baccalauréat de Mathématiques soit supérieure à 10.

6.4 Distributions d'échantillonnage de la variance S^2 :

Soit une population, et soit X V.A telque $E(X) = m$ et $var(X) = \sigma^2$, et un échantillon de taille n, Et on sait que la moyenne de l'échantillon est

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Alors puisque $E(X) = m$ et $var(X) = \sigma^2$, on a :

1. **La moyenne de S^2 :**

$$E(S^2) = \frac{n-1}{n}\sigma^2.$$

2. **La variance de S^2 :**

$$\begin{aligned} Var(S^2) &= 2(n-1)\frac{\sigma^4}{n^2} \\ \sigma_{S^2} &= \sqrt{2(n-1)}\frac{\sigma^2}{n} \end{aligned}$$

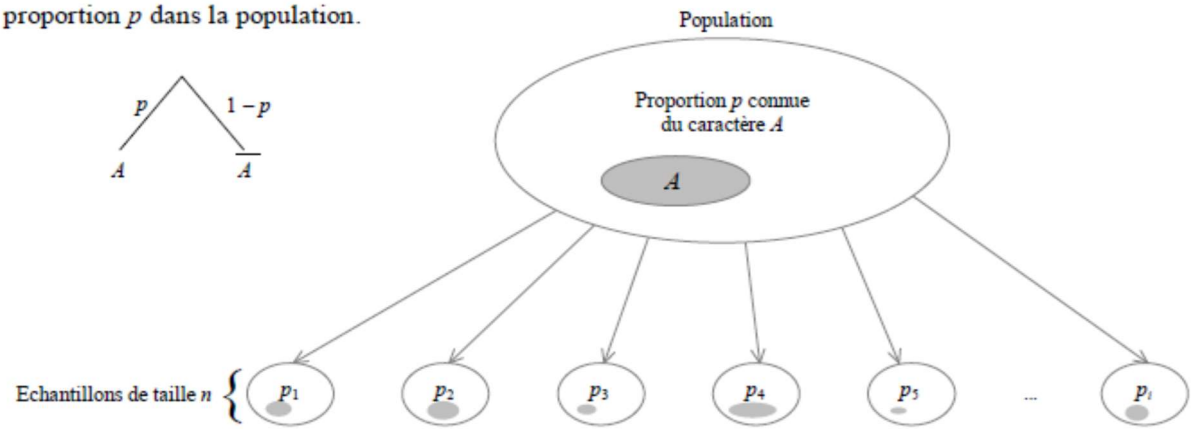
3. **La loi de $n\frac{S^2}{\sigma^2}$:**
   Si $X \rightsquigarrow N(m, \sigma)$ Alors

$$n\frac{S^2}{\sigma^2} \rightsquigarrow \chi_{n-1}$$

χ_{n-1} :(Khi-deux $n-1$ degret de libérté)

6.5 Distributions d'échantillonnage de la Proportion F:

Cette fois-ci, on dispose d'une population sur laquelle on étudie un caractère (ou attribut) A dont on connaît la proportion p dans la population.



On s'intéresse aux échantillons de taille n. La proportion du caractère A dans les échantillons sera-t-elle toujours la même ? Evidemment non, cette proportion varie en fonction de l'échantillon choisi. Notons F la variable aléatoire qui, à chaque échantillon de taille n, associe sa proportion du caractère A (F s'appelle *distribution des fréquence des échantillons*). Que peut-on dire de cette variable aléatoire F ?

Soit une population, qui se dévise sur deux parties : A et $\overline{A}$ tel que $P(A) = p$ et $P(\overline{A}) = 1 - p$, et un échantillon de taille n.

Soit K_n le nombre des indévidus qui ont l'evenement A.

$$F = \frac{K_n}{n}$$

Alors puisque $E(X) = m$ et $var(X) = \sigma^2$, on a :

1. **La moyenne de F :**
$$E(F) = p.$$
2. **La variance de F :**
$$Var(F) = \frac{(p-n)}{n}p$$
$$\sigma_F = \sqrt{\frac{(p-n)}{n}p}$$
3. **La loi de F :**
Si $K_n \rightsquigarrow B(n,p)$ Alors
$$F \rightsquigarrow N(p; \sqrt{\frac{(p-n)}{n}p})$$

Démonstration :

Nous allons avoir ici un modèle binomial ou apparenté dont on sait qu'il converge vers la loi normale.

Pour tout i compris entre 1 et n, notons X_i la variable aléatoire définie par :

$$X_i = \begin{cases} 1 \text{ si le } i\text{-ème élément de l'échantillon possède l'attribut } A \\ 0 \text{ sinon} \end{cases}$$

La variable aléatoire X_i suit une loi de Bernoulli de paramètre p.

La variable aléatoire $X = X_1 + X_2 + ... + X_n$ est donc binomiale de paramètres n et p :

$$X \rightsquigarrow B(n, p)$$

En conséquence : $E(X) = np$ et $\sigma(X) = \sqrt{np(1-p)}$

La variable aléatoire $F = \frac{X}{n}$ correpond ainsi à la fréquence de l'attribut A dans l'échantillon.

D'après les propriétés de l'espérance et de l'écart-type :

$$E(F) = \frac{E(X)}{n} = p \text{ et } \sigma(F) = \frac{\sigma(X)}{n} = \sqrt{\frac{p(1-p)}{n}}$$

Exemple :

Une élection a eu lieu et un candidat a eu 40 % des voix.

On prélève un échantillon de 100 bulletins de vote.

Quelle est la probabilité que, dans l'échantillon, le candidat ait entre 35 % et 45 % des voix ?

Ici, nous avons $n = 100$ et $p = 0,4$. La variable aléatoire F correspondant à la fréquence des votes pour le candidat dans l'échantillon vérifie donc :

$$F \rightsquigarrow N\left(0,4;\sqrt{\frac{0,4\times 0,6}{100}}\right) = N\left(0,4;\frac{\sqrt{0,24}}{10}\right)$$

Posons $T = \dfrac{F-0,4}{\frac{\sqrt{0,24}}{10}}$ ainsi $T \rightsquigarrow N(0\ ;\ 1)$. Nous obtenons alors par centrage et réduction :

$$P(0,35 \leqslant F \leqslant 0,45) = P(-1,02 \leqslant T \leqslant 1,02) = 2\Pi(1,02) - 1$$

Et par lecture directe de la table de la loi normale centrée-réduite $\Pi(1,02) = 0,8461$.

D'où : $$P(0,35 \leqslant F \leqslant 0,45) = 0,6922$$

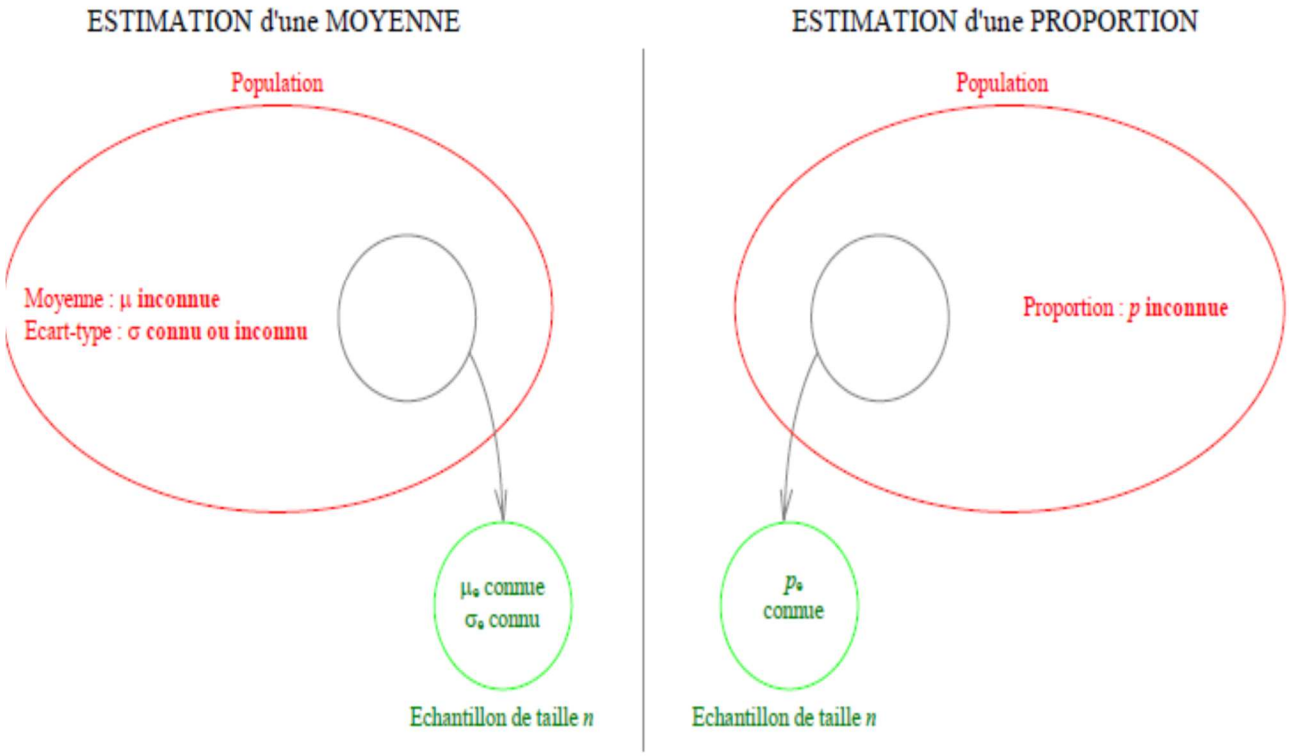
Il y a donc environ 69 % de chance que, dans un échantillon de taille $n = 100$, le candidat ait entre 35 % et 45 % des voix.

En analysant l'exercice ci-dessus, on constate que l'on dispose des informations sur la population (ici, l'ensemble des votes) parce que l'élection a déjà eu lieu. On en déduit des informations sur l'échantillon. Mais, dans la pratique, c'est souvent le phénomène réciproque que nous étudierons : les élections n'ont pas encore eu lieu et on voudrait retrouver les informations sur la population grâce un sondage réalisé sur un échantillon. D'où la partie suivante de ce document consacrée à l'estimation.

7. CHAPITRE 6 : THÉORIE DE L'ESTIMATION

L'objectif de cette partie est de répondre à la problématique suivante : *comment, à partir d'informations (couple moyenne/écart-type ou proportion) calculées sur un échantillon, retrouver ou plutôt estimer celles d'une population entière ?* L'estimation est le problème réciproque de l'échantillonnage. (Mais nous aurons besoin des résultats établis sur la théorie de l'échantillonnage pour passer à la phase estimative).

Nous distinguerons deux cas : celui où l'on cherche à estimer la moyenne μ d'une variable aléatoire définie sur une population et celui où l'on cherche à estimer la proportion d'individus *p* ayant tel caractère dans la population.



7.1 Estimation ponctuelle

7.1.1 Estimation de la moyenne

Contexte : on considère une variable aléatoire X sur une population de moyenne (ou espérance) μ inconnue et d'écart-type σ inconnu (ou connu). On suppose que l'on a prélevé un échantillon de taille n (tirage avec remise ou assimilé) sur lequel on a calculé la moyenne μ_e et l'écart-type σ_e.

Une estimation ponctuelle $\widehat{\mu}$ de la moyenne μ de la population est :

$$\widehat{\mu} = \mu_e$$

Une estimation ponctuelle $\widehat{\sigma}$ de l'écart-type σ_e de la population est :

$$\widehat{\sigma} = \sqrt{\frac{n}{n-1}}\,\sigma_e$$

Le coefficient $\sqrt{\frac{n}{n-1}}$ s'appelle *correction de biais*. Lorsque la taille n de l'échantillon est assez grand (en pratique $n \geqslant 30$), ce coefficient est très voisin de 1, si bien que, dans ce cas, on peut estimer $\widehat{\sigma} \simeq \sigma_e$.

Exemple :

Une université comporte 1500 étudiants. On mesure la taille de 20 d'entre eux. La moyenne μ_e et l'écart-type σ_e calculés à partir de cet échantillon sont :

$$\mu_e = 176 \text{ cm} \quad \text{et} \quad \sigma_e = 6 \text{ cm}$$

Nous pouvons donc estimer les paramètres de la population :

$$\widehat{\mu} = 176 \text{ cm} \quad \text{et} \quad \widehat{\sigma} = \sqrt{\frac{20}{19}} \times 6 \simeq 6{,}16 \text{ cm}$$

7.1.2 Estimation de la variance :

La variance observée constitue le meilleur estimateur de σ^2 , variance de la loi de probabilité de la variance aléatoire X lorsque l'espérance μ est connue :

$$\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Xi} - \boldsymbol{\mu})^{\boldsymbol{2}}$$

Remarque : cette estimation de la variance de la population est rarement utilisée dans la mesure où n'est pas connue, l'espérance μ ne l'est pas non plus

Le meilleur estimateur de $\hat{\sigma}^2$, variace de la loi de probabilité de la variable aléatoire X lorsque l'espérance μ est inconnue est :

$$\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}} = \frac{\boldsymbol{n}}{\boldsymbol{n-1}} \boldsymbol{S}^{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Xi} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}$$

Remarque : lorsque n augmente, la variance observée S^2tend vers la variance de la population σ^2

$$\lim_{\boldsymbol{n \leftarrow +\infty}} \boldsymbol{S}^{\boldsymbol{2}} = \lim_{\boldsymbol{n \leftarrow +\infty}} \frac{(\boldsymbol{n-1})}{\boldsymbol{n}} \boldsymbol{\sigma}^{\boldsymbol{2}} = \boldsymbol{\sigma}^{\boldsymbol{2}}$$

7.1.3 Estimation de la proportion

Soit le schéma de Bernoulli dans lequel le caractère A correspond au succès. On note *p* la fréquence des individus de la population possédant le caractère A. La valeur de ce paramètre étant inconnu, on cherche à estimer la fréquence *p* à partir des données observables sur un échantillon. A chaque échantillon non exhaustif de taille *n.* on associe l'entier *k.* nombre d'individus possédant le caractère A.

Soit *K* une variable aléatoire discrète suivant une loi binomiale *B(n.p)* et pour laquelle on souhaite estimer la fréquence *p.* La fréquence observée du nombre de succès observé dans un échantillon de taille *n* constitue le meilleur estimateur de *p :*

$$\widehat{\boldsymbol{P}} = \frac{\boldsymbol{K}}{\boldsymbol{n}}$$

Exemple :

On a prélevé au hasard, dans une population de lapin, 100 individus. Sur ces 100 lapins, 20 sont atteints par la myxomatose. Le pourcentage de lapins atteints par la myxomatose dans la population est donc :

$$\hat{p} = \frac{K}{n} = \frac{20}{100} = 0{,}2 \text{ soit } 20\% \text{ de lapins atteins dans la population}$$

Remarque :

Nous n'avons fait qu'une estimation, il est bien sûr impossible de retrouver les vraies caractéristiques μ et σ de la population.

L'estimation ponctuelle permet surtout de disposer d'une valeur de référence pour poursuivre/affiner les calculs.

On souhaiterait notamment pouvoir faire une estimation par intervalle, en contrôlant le risque pris.

7.2 Estimation par intervalle de confiance

7.2.1 Intervalle de confiance d'une moyenne

L'intervalle de confiance de la moyenne μ pour un coefficient de risque α est donc

$$\bar{X} - \varepsilon_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \varepsilon_\alpha \frac{\sigma}{\sqrt{n}}$$

Quelque soit la valeur de n si X→ $N(\mu, \sigma)$ et la variance σ^2 *est connue*

L'intervalle de confiance de la moyenne μ pour un coefficient de risque α est donc

$$\bar{X} - t_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + t_\alpha \frac{\sigma}{\sqrt{n}}$$

Quelque soit la valeur de n si X→ $N(\mu, \sigma)$ et la variance σ^2 *est inconnue*

Remarque : lorsque n>30, la loi de Student converge vers une loi normale reduite. Ainsi la valeur de t_α (n-1) est égale à ε_α

Exemples :

(1) Dans un échantillon de **20 étudiants** de même classe d'âge et de même sexe, la taille moyenne observée est de 1,73m et l'écart-type de 10 cm. La taille moyenne de l'ensemble des étudiants de l'université est donc :

avec $\bar{x} = 1{,}73\text{m}$; $\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{20}{19} \times 0{,}01 = 0{,}011$ et $t_\alpha = 2{,}086$

d'où $t_\alpha \frac{\hat{\sigma}}{\sqrt{n}} = 2{,}086 \times \sqrt{\frac{0{,}011}{20}} = 0{,}049$ ainsi $\mu = \bar{X} \pm t_\alpha \frac{\hat{\sigma}}{\sqrt{n}} = \mathbf{1{,}73\text{m} \pm 0{,}049}$

La **taille moyenne** des étudiants dans la population est comprise dans l'intervalle **[1,68 ; 1,78]** avec une probabilité de 0,95.

(2) Dans un échantillon de **100 étudiants**, la taille moyenne de la population est :

$\bar{x} = 1{,}73\text{m}$; $\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{100}{99} \times 0{,}01 = 0{,}01$ et $\varepsilon_\alpha = 1{,}960$

d'où $\varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}} = 1{,}960 \times \sqrt{\frac{0{,}010}{100}} = 0{,}02$ ainsi $\mu = \bar{X} \pm \varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}} = \mathbf{1{,}73\text{m} \pm 0{,}02}$

La **taille moyenne** des étudiants dans la population est comprise dans l'intervalle **[1,71 ; 1,75]** avec une probabilité de 0,95.

Ainsi lorsque **la taille** de l'échantillon **augmente** pour un même coefficient de confiance $(1-\alpha)$, l'estimation autour de μ est **plus précise**.

- **Si $n > 30$ et X suit une loi inconnue,**

La démarche est la même que pour le cas précédent puisque par définition la variance de la population est inconnue et doit être estimée avec la variance observée :

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2$$ (voir **estimation ponctuelle**)

Comme pour le cas 1, la loi suivie par la variable centrée réduite $\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \rightarrow \mathcal{N}(0,1)$ (**conditions**).

L'intervalle de confiance de la moyenne μ pour un coefficient de risque α est donc

$$\bar{X} - \varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + \varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$$

Vrai seulement si n est grand

Si n<30 et X suit une loi inconnue, la loi de probabilité n'est pas connue et l'on a recours aux statistiques non paramétriques

8.2.2 Intervalle de confiance d'une proportion

L'intervalle de confiance de la fréquence P pour un coefficient de risque α est donc

$$\frac{k}{n} - \varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \frac{k}{n} + \varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Vrai seulement si n est grand et np, nq >5

Remarque : si la taille de l'échantillon est faible on a recours aux lois exactes

Exemple :

Un laboratoire d'agronomie a effectué une étude sur le maintien du pouvoir germinatif des graines de *Papivorus subquaticus* après une conservation de 3 ans.
Sur un lot de 80 graines, 47 ont germé. Ainsi la probabilité de germination des graines de *Papivorus subquaticus* après trois ans de conservation avec un coefficient de confiance de 95% est donc :

avec $\hat{p} = \frac{K}{n} = \frac{47}{80} = 0{,}588$, $\hat{q} = \frac{n-K}{n} = \frac{33}{80} = 0{,}412$ et $\varepsilon_\alpha = 1{,}96$

alors $\varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1{,}96 \times \sqrt{\frac{0{,}588 \times 0{,}412}{80}} = 0{,}108$ d'où $p = 0{,}588 \pm 0{,}108$

ainsi **la probabilité de germination est** comprise dans l'intervalle **[0,480 et 0,696]** avec une probabilité de **0,95.**

8.2.3 Intervalle de confiance d'une variance :

Si on s'intéresse à la variance σ^2 d'une population normale, l'estimation par intervalle de confiance consiste à déterminer les bornes σ^2_1 et σ^2_2 d'un intervalle qui a un niveau de confiance (1-α) de contenir σ^2.

Les limites σ^2_1 et σ^2_2 sont telles que :

$$p(\sigma^2_1 \leq \sigma^2 \leq \sigma^2_2) = 1 - \alpha$$

Les limites de confiances sont alors

$$\sigma^2_1 = \frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \text{et} \quad \sigma^2_2 = \frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{\chi^2_{\frac{\alpha}{2}}}$$

Les valeurs de $\chi^2_{\frac{\alpha}{2}}$ et $\chi^2_{1-\frac{\alpha}{2}}$ sont à (n-1) degré de liberté.

8. CHAPITRE 7 : LES TESTS D'HYPOTHÈSES

8.1 Introduction :

Un test d'hypothèse est un procédé d'inférence statistique permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations. Les méthodes de l'inférence statistique nous permettent de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

On distinguera deux classes de tests :

- **Les tests paramétriques** requirent un modèle à fortes contraintes (normalité des distributions ou approximation normale pour des grands échantillons).
- **Les tests non paramétriques** sont des tests dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. Il n'y a pas d'hypothèse de normalité au préalable.

Pour **les tests paramétriques** nous distinguons :

- **Le test de conformité** consiste à confronter un paramètre calculé sur l'échantillon à une valeur pré-établie. Les plus connus sont certainement les tests portant sur la moyenne, la variance ou sur les proportions. Ils sont destinés à comparer entre eux une population théorique et un échantillon observé. Ils servent à vérifier si un échantillon donné peut être considéré comme extrait d'une population possédant telle caractéristique particulière (telle moyenne, telle variance, …). Le test se fait en vérifiant si la différence entre la valeur observée et la valeur théorique du paramètre considéré peut être attribuée au hasard ou non.
- **Les tests d'homogénéité ou d'égalité** ont pour but de comparer entre elles un certain Nombre de populations, à l'aide d'un même nombre d'échantillons.

8.2 Principe de test

Le principe des tests d'hypothèse est de poser une hypothèse de travail et de prédire les conséquences de cette hypothèse pour la population ou l'échantillon. On compare ces prédictions avec les observations et l'on conclut en acceptant ou en rejetant l'hypothèse de travail à partir de règles de décisions objectives.

Différentes étapes doivent être suivies pour tester une hypothèse :

- Définir l'hypothèse nulle, notée H0, à contrôler

- Calculer à partir de l'échantillon les paramètres statistiques à tester
- Choisir le test statistique approprié et sa loi de distribution
- Calculer, à partir des données fournies par l'échantillon, la valeur de la statistique du test
- Définir le niveau de signification du test et la région critique associée
- Comparer les valeurs (observée et critique) et prendre une décision concernant l'hypothèse posée.

L'hypothèse nulle notée H0 est l'hypothèse que l'on désire contrôler : elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et est due aux fluctuations d'échantillonnage. Cette hypothèse est formulée dans le but d'être rejetée.

L'hypothèse alternative notée H1 est la "négation" de H0, elle est équivalente à dire « H0 est fausse ». La décision de rejeter H0 signifie que H1 est réalisée ou H1 est vraie.

8.3 Risque d'erreur

Il existe deux stratégies pour prendre une décision en ce qui concerne un test d'hypothèse : la première stratégie fixe à priori la valeur du seuil de signification α et la seconde établit la valeur de la probabilité critique α obs à posteriori.

On appelle risque d'erreur de première espèce α la probabilité de rejeter H0 et d'accepter H1 alors que H0 est vraie. Cela signifie de déclarer une différence qu'en réalité elle est nulle et est due uniquement aux fluctuations d'échantillonnage.

8.4 Puissance d'un test

Rappelons que les tests ne sont pas faits pour « démontrer » H0 mais pour « rejeter » H0. L'aptitude d'un test à rejeter H0 alors qu'elle est fausse constitue la puissance du test. On appelle puissance d'un test, la probabilité de rejeter H0 et d'accepter H1 alors que H1 est vraie. Sa valeur est 1 – β.

	Réalité		
Décision		H0 Vraie	H0 Fausse
	H0 Acceptée	Correct	Manque de puissance Risque β
	H0 Rejetée	Rejet à tort Risque α	Puissance du test 1-β

8.5 Les principaux tests statistiques utilisés

8.5.1 Test de conformité (Cas d'un seul échantillon)

a- Test relatif aux proportions

Le modèle mathématique est le suivant. On dispose d'une population dans laquelle chaque individu présente ou non un certain caractère, la proportion d'individus présentant le caractère étant notée p, et un échantillon aléatoire de taille n extrait de cette population. La proportion f calculée à partir de l'échantillon est considérée comme une réalisation d'une VA de loi binomiale B(n; p) qu'on peut assimiler, si n est assez grand, à une loi normale N(p, $\sqrt{\frac{p(1-p)}{n}}$) On veut tester H0 : p = p0 contre H1 : p ≠ p0.

Le test utilisé est celui de Z, $Z_{ob} = \frac{P-P0}{\sqrt{\frac{P0(1-P0)}{n}}}$

H0 est rejetée (Différence statistiquement significative) si Zob > Zth (Zth = 1,96 ; 2,58 et 3,20 à respectivement α = 0,05 ; 0,01 et 0,001).

b- Test relatif aux moyennes

Cas de Grand échantillon ou variance connue :

On suppose que l'on a un échantillon qui suit une loi normale $N(\mu,\sigma^2)$ ou la variance est connue. On veut tester H0 : $\mu = \mu_0$ contre H1 : $\mu \neq \mu_0$, le test utilisé est celui de Z

$Z_{ob} = \frac{m-\mu 0}{\frac{\sigma}{\sqrt{n}}}$ suit une loi Normale centré réduite *N(O ;1)*

Ho est rejetée (Différence statistiquement significative) si $Z_{ob} > Z_{th}$ (Z_{th} = 1,96; 2,58 et 3 ,20 à respectivement α=0.05 ; 0.01 ; 0.001)

Cas de petits échantillons (variance inconnue):

On suppose que l'on a un échantillon qui suit une loi normale $N(\mu,\sigma^2)$ ou la variance est maintenant inconnue. On veut tester H0 : $\mu = \mu_0$ contre H1 : $\mu \neq \mu_0$, Comme la variance est inconnue, on l'estime par la variance empirique de l'échantillon. La statistique du test $T_{ob} = \frac{m-\mu_0}{\frac{S}{\sqrt{n}}}$ suit une loi de Student à (n-l) degré de liberté.

H_n est rejetée (Différence statistiquement significative) si $t_{ob} > t_{th}$ (extrait de la table de Student à α taux d'erreur et (n-1) ddl).

8.5.2 Tests d'égalité (Cas d'échantillons indépendants)

a- Test relatif aux proportions

On veut comparer deux proportions p1 et p2 à partir de deux échantillons. Le modèle mathématique est le suivant. On considère les proportions p1 et p2 associés aux deux échantillons. On veut tester H0 : p1 = p2 contre H1 : p1 6= p2. On prend la statistique

On prend la statistique $Z_{ob} = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)(\frac{1}{n_1}+\frac{1}{n_2})}}$ *Avec* $p_0 = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

H_0 est rejetée (Différence statistiquement significative) si $Z_{ob} > Z_{th}$ (Z_{th} = 1,96 ; 2,58 et 3,20 à respectivement α = 0,05 ; 0,01 et 0,001).

b- Test relatif aux moyennes

Cas de grand échantillon

On suppose que l'on a deux échantillons (X1; : : : ;Xn1) et (Y1; : : : ; Yn2) qui suivent une loi normale $N(\mu 1; \sigma_1^2)$ et $N(\mu 2; \sigma_2^2)$.

On veut tester H0 : $\mu_1 = \mu_2$ contre H1 : $\mu_1 \neq \mu_2$,

La statistique $Z_{ob} = \frac{m_1 - m_2}{\sqrt{(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2})}}$ suit une loi normale centré réduite

H_0 est rejetée (Différence statistiquement significative) si $Z_{ob} > Z_{th}$ (Z_{th} = 1,96 ; 2,58 et 3,20 à respectivement α = 0,05 ; 0,01 et 0,001)

Cas de petits échantillons avec variances égales

On veut tester H0 : $\mu_1 = \mu_2$ contre H1 : $\mu_1 \neq \mu_2$, on calcule au préalable la variance commune estimée pour les deux échantillons $S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

Alors la variable aléatoire $T_{ob} = \frac{m_1 - m_2}{S\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$ suit une loi de Student à ($n_1 + n_2 - 2$) degré de liberté

H_0 est rejetée si $t_{ob} > t_{th}$ (à α taux d'erreur et (n1 + n2 – 2) ddl).

Cas de petits échantillons avec variances inégales

On veut tester H0 : $\mu_1 = \mu_2$ contre H1 : $\mu_1 \neq \mu_2$, on calcule au préalable la variance commune estimée pour les deux échantillons $S^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}$

Alors la variable aléatoire $T_{ob} = \frac{m_1-m_2}{S\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$ suit une loi de Student à (ν) degré de liberté

Où (ν) est l'entier le plus proche de la correction de Welch.

$$ddl_{Welch} = \frac{\left(\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}\right)^2}{\frac{\left(\frac{{S_1}^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{{S_2}^2}{n_2}\right)^2}{n_2-1}}$$

H_0 est rejetée si $t_{ob} > t_{th}$ (à α taux d'erreur et (ν) ddl).

8.5.3 Tests d'égalité (Cas d'échantillons appariés) :

Les échantillons sont dites appariés ou liés si la ième observation de l'échantillon 1 dépend de la ième observation de l'échantillon 2. C.-à-d. quand on fait deux mesures de même nature (par exemple avant et après un traitement) sur les mêmes sujets.

Dans ce cas, il ne faut pas procéder comme si les deux échantillons étaient indépendants et vouloir comparer la moyenne des xi à la moyenne des yi. Ce serait méconnaître le problème posé ; les deux mesures sont faites sur le même sujet, il y a donc un effet « sujet ».

Il convient de comparer chaque xi à l'yi correspondant et pour cela il faut effectuer les n différences : $d_i = x_i - y_i$

On obtient donc un échantillon unique de n valeurs d_i, sur lequel on calcule la moyenne $\bar{d}$ et l'écart type S_d. Si les x_i ne sont pas différents des y_i, alors la moyenne $\bar{d}$ ne doit pas différer beaucoup de zéro. $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$; $S_d = \sqrt{\frac{\sum(d_i-\bar{d})^2}{n-1}}$

On est donc ramené au problème précédent de la comparaison d'une moyenne observée $\bar{d}$ à une valeur théorique 0.

On veut tester H0 : $\bar{d} = 0$ contre H1 : $\bar{d} \neq 0$,

Le test utilisé est celui de t (Student) réalisé sur les écarts des valeurs observées.

Alors la variable aléatoire $T = \frac{\bar{d}-0}{\frac{S_d}{\sqrt{n}}}$ suit une loi normale à (n-1) degré de liberté.

H0 est rejetée si $T_{ob} > T_{th}$ (à α taux erreur et (n – 1) ddl) ; n est le nombre de pair d'observations.

8.5.4 Test d'indépendance

Les tests d'indépendance ont pour but de contrôler l'indépendance stochastique de deux ou plusieurs critères de classification. Ils permettent également d'effectuer des comparaisons de proportions.
La distribution de probabilité correspondante est une distribution à deux dimensions, et les données relatives à tout échantillon sont présentées sous la forme d'un tableau de contingence.
Pour des échantillons aléatoires et simples, si les deux critères de classification sont indépendants, les probabilités pij de la distribution à deux dimensions peuvent être estimées par :

$$\widehat{\boldsymbol{Pij}} = \boldsymbol{fi} \times \boldsymbol{fj} \text{ avec } \boldsymbol{fi.} = \frac{\boldsymbol{ni.}}{\boldsymbol{n}} \text{ et } \boldsymbol{f.j} = \frac{\boldsymbol{n.j}}{\boldsymbol{n}}$$

ni. et n.j sont les effectifs marginaux, et nij les effectifs conjoints.
Les effectifs attendus correspondants sont donc :

$$\mathrm{n}\,\hat{p}_{ij} = \mathrm{n}\, f_{i.} \times f_{.j} = \mathrm{n} \frac{n_{i.}}{n} \frac{n_{.j}}{n} = \frac{n_{i.} \times n_{.j}}{n}$$

Les effectifs attendus doivent tous être supérieurs ou égales à 5.

Formulation de l'hypothèse nulle :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés nij et les effectifs attendus ou théoriques correspondants n *pij*

L'hypothèse nulle est l'indépendance des deux critères de classification

H0 : nij=n$\widehat{p}$ij

les effectifs attendus doivent touts être supérieurs ou égales à 5.

Variable de décision :

La comparaison des effectifs observés et attendus se fait comme pour les tests d'ajustement, en calculant la variable de décision suivante :

$$\mathrm{VD} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{nij^2}{\hat{n p}_{ij}} - \mathrm{n}$$

On démontre que la variable de décision est une variable aléatoire Khi deux avec (p-1)(q-1) degré de liberté.

Région d'acceptation :

La valeur critique qui délimite la région d'acceptation est x^2 telle que :

$p(VD < \chi^2) = 1 - \alpha \rightarrow \chi^2 = \chi^2_{1-\alpha}$

Le test étant toujours unilatéral, la région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

On rejettera donc l'hypothèse nulle lorsque la valeur de la variable de décision est supérieure ou égale à $\chi^2_{1-\alpha}$

BIBLIOGRAPHIE

- DJOUDAD-KADJI HAFSA., 2016-2017 : La statistique descriptive univariée appliquée à la biologie . Polycopié de cours, Université de Bejaia- Algerie.

- ADIL ELMARHOUM., 2013 : ECHANTILLONNAGE ET ESTIMATIONS, Polycopié de cours, Université Mohamed v – Agdal- Maroc

- COSTANTINI GILLE, Echantillonnage-estimation, statistiques inférentielles, BTS 2ème année , p1 https://docplayer.fr/13457270-Partie-a-echantillonnage.html

- MOHAMED BARRADI ., 2014 : Résumé sur L'échantillonnage . Résumé-1-echantiol-s3 (studylibfr.com)

- PAUL MILAN,. 2015 : Lois de probabilité à densité Loi normale 10_cours_proba_cond_loi_binomiale.pdf (lyceedadultes.fr)

- MOUFFOK ., Biostatistique et analyse des données , Université de Setif, M1 PA 19-20 Biostatistiques Mouffok.pdf (univ-setif.dz)

- MONDHER ABROUGUI **.,** 2008 : Biostatistique-1-Cours & Activités **,** institut supérieur de l'éducation et de la formation continue-Tunis- Tunisie.

- JEAN VAILLANT., 2015 : Eléments de Statistique descriptive, Université des Antilles, elémentsstatdes_avec_exos.pdf (weebly.com), France.

- N. BENDIMRAD.,2018/2019 : Dénombrement et analyse combinatoire, Université Oran1-Algerie.

- CHARLES SUQUET.,2002/2003 : Introduction au Calcul des Probabilités , Université des

- Sciences et Technologies de Lille U.F.R. de Mathématiques Pures et Appliquées .

- **Site de Marcel Délèze.,** Le théorème de Bayes, démonstration et exemple (deleze.name) ? https://www.deleze.name/marcel/